THE UNIVERSITY OF ALBERTA

METHODS FOR AUTOMATIC DIAGNOSIS OF DISEASE

by

© JOHN THOMAS CUMBERBATCH

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

FALL, 1974

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled METHODS FOR AUTOMATIC DIAGNOSIS OF DISEASE, submitted by John Thomas Cumberbatch in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Henna
Supervisor

Z.K. Schubert

Kellogg Wilson

Michael Grace

Stan Cabay

Perry Schein
External Examiner

Date: Sep-4-74.

ABSTRACT

Three methods for automatic diagnosis of disease are formulated and applied to a data base of several hundred gastroenterological patients, each determined, by radiological diagnoses, as having one of six diseases. Application of these methods requires no assumptions regarding statistical independence of the symptoms. Any order of dependence between the symptoms and each disease may be allowed for by appropriate choice of terms in disease-symptom functions.

The first method diagnosis each patient as having the disease corresponding to that disease-symptom function having the largest value, as evaluated from the patient's symptoms. Parameters may be used to change the coefficients of the disease-symptom functions linearly and non-linearly in order to obtain a maximum number of correct diagnoses of patients in the data base. The method is used to determine the disease of patients not contained in the data base. The resulting accuracy of diagnosis is discussed with regard to the size of the data base and the effect of inclusion of non-linear and interactive terms in the disease-symptom functions.

The second method uses the values of the disease-symptom functions, as evaluated from each patient's symptoms, to determine the probability that each patient

has each disease. A constrained solution is found to the problem of determining the coefficients of the disease-symptom functions in order that the disease probabilities result in the correct diagnosis of a maximum number of patients in the data base. The method is used to determine the most probable disease of each patient in the data base.

The accuracy of diagnosis which results from using these two methods is compared with that obtained using other methods for automatic diagnosis. Reasons are given as to why the methods formulated herein produce superior results.

The third method is one of sequential diagnosis in which additional symptoms are chosen according to their diagnostic value per unit cost. The diagnostic value of each symptom is an indirect measure of the increase in accuracy of diagnosis that results from the addition of that symptom. This diagnostic value is disease conscious, being formulated in terms of the expression used to determine the coefficients of each disease-symptom function. The method is used to diagnose patients not contained in the data base. A comparison of a disease conscious and a non-disease-conscious selection of additional symptoms shows that the former can lead to a definitive diagnosis using fewer symptoms than the latter.

ACKNOWLEDGEMENTS

I express my appreciation to Dr. D. Fenna, my supervisor, for his advice, guidance and criticism through the course of this research. I am indebted to Prof. H.S. Heaps and Dr. K.V. Leung for their assistance in the initial stage of the research. Thanks are due to Dr. M. Grace and Dr. L. Schubert for their suggestions and criticisms. Comments from Dr. S. Cabay and Dr. K. Wilson are also acknowledged. I am grateful to Mrs. Mary Yiu for typing the manuscript.

The financial assistance provided by the National Research Council of Canada, in the form of a scholarship, and the Department of Computing Science, in the form of teaching assistantships, is appreciated.

Finally I express great appreciation to my wife, Mona, whose encouragement and many hours of proof reading have made the preparation of this thesis so much easier.

CONTENTS

		Page
CHAPTER 1	INTRODUCTION	1
CHAPTER 2	LITERATURE SURVEY	4
2.1	Introduction	4
2.2	Patients, Symptoms, Diseases	6
2.3	Bayesian Methods	8
2.3.1	The Work of Reale	11
2.3.2	The Work of Warner	12
2.3.3	The Work of Boyle	13
2.3.4	The Work of Scheinok	14
2.3.5	Comment	16
2.4	Corrections to Bayes' Theorem	16
2.4.1	Vanderplas' Correction	17
2.4.2	Best's Correction	18
2.4.3	Discussion	18
2.5	Non-Bayesian Methods	19
2.5.1	Boolean Algebra	19
2.5.2	Weighted Symptom Summation	20
2.5.3	Discriminant Analysis	20
2.5.4	Least Squares Fit	21
2.5.5	K-Nearest-Neighbours Rule	22
2.5.6	Discussion	23
2.6	Allowance for Dependence Between Symptoms	24
2.6.1	A Bayesian Method	24
2.6.2	A Non-Bayesian Method	25
2.6.3	Discussion	25
2.7	Conclusion	26

		Page
CHAPTER 3	DISEASE-SYMPTOM FUNCTIONS	29
3.1	Introduction	29
3.2	Disease-Symptom Functions	31
3.3	Linear Disease-Symptom Functions	38
3.4	Symptom Quantization	41
3.5	Generalized Linear Disease-Symptom Functions	42
3.6	A First Estimate for Alpha	45
3.7	Confidence Limits	46
3.8	Results Using Disease-Symptom Functions	47
3.8.1	Diagnosis of Previous Patients	47
3.8.2	Diagnosis of New Symptom Vectors	52
3.9	Conclusion	57
CHAPTER 4	DISEASE PROBABILITIES	58
4.1	Introduction	58
4.2	The Assumption of Normality	60
4.3	Disease Probabilities	61
4.3.1	Limited Disease Probabilities	62
4.3.2	Formulations for Suitable Coefficients	66
4.3.3	Relation to the Alpha-Beta Method	68
4.4	Extensions to Disease Probabilities	71
4.4.1	Extended Disease Probabilities	72
4.4.2	Further Extended Disease Probabilities	74
4.4.3	Advantages	76
4.4.4	Determining Suitable Coefficients	78

		Page
CHAPTER 4	(cont'd)	
4.5	Results Using Disease Probabilities	83
4.5.1	Assuming Normality	84
4.5.2	Non-Normality	88
4.5.3	Comment	94
CHAPTER 5	COMPARISON OF RESULTS	97
5.1	Introduction	97
5.2	Linear Separating Surfaces	97
5.3	Non-Linear Separating Surfaces	101
CHAPTER 6	SEQUENTIAL DIAGNOSIS	103
6.1	Introduction	103
6.2	The Diagnostic Value of a Symptom	107
6.2.1	With Respect to One Disease	108
6.2.2	With Respect to Several Diseases	114
6.3	The Diagnostic Value of Several Symptoms	115
6.4	Results Using Sequential Diagnosis	117
6.4.1	Symptom Sequences	118
6.4.2	The Diagnosis of New Patients	120
6.4.3	An On-line Interactive System	127
6.5	Conclusion	131
CHAPTER 7	SUMMARY, CONCLUSIONS AND RECOMMENDATIONS	132
BIBLIOGRAPHY		140

Page

APPENDIX 1 146

APPENDIX 2 148

APPENDIX 3 150

APPENDIX 4 153

LIST OF TABLES

Table		Page
3.1	Number of Previous Patients Correctly Diagnosed When Using Linear Disease-Symptom Functions	49
3.2	Number of Symptom Vectors Correctly Associated with Their Disease When Using Linear Disease-Symptom Functions	51
4.1	Second Estimates of β_k to the Nearest Multiple of 0.2	85
4.2	Values of β_k found to Maximize J_k to the Nearest Multiple of 0.2	85
4.3	The Value of J_k for Different Values of β_k	86
4.4	Percentage Accuracy of Diagnosis of 223 Previous Patients Using Disease Probabilities Assuming Normality	87
4.5	Frequency Distribution Analysis of z^{kl} ($k = 5,6$)	89
4.6	Percentage Accuracy of Diagnosis of 223 Previous Patients Using Disease Probabilities Not Assuming Normality	93
4.7	Number of Previous Patients Correctly Diagnosed When Using Further Extended Disease Probabilities Not Assuming Normality	95
4.8	Number of Symptom Vectors Correctly Associated With Their Disease When Using Further Extended Disease Probabilities Not Assuming Normality	96
6.1	Disease-Conscious and Non-Disease-Conscious Symptom Sequences	119
6.2	Values of $P(D_k \mid Z_k(n^*))$ for Three New Patients Using Six and 11 Symptoms	122

Table		Page
6.3	Average Values of $P(D_k \mid Z_k(n^*))$ for Sequential Diagnosis of New Patients Known to Have D_k	124
6.4	Probability Difference Table	126

LIST OF FIGURES

Figure		Page
3.1	Linear Separation of Two Disease Sets	32
3.2	Linear Separating Surfaces for Three Disease Sets	32
3.3	Percentage Accuracy of Diagnosis of Previous Symptom Vectors	53
3.4	Percentage Accuracy of Diagnosis of New Symptom Vectors	53
4.1	Histogram of Transformed $Z_k(n)$ for Patients Having Cancer	92
4.2	Histogram of Transformed $Z_k(n)$ for Patients Having Gallstones	92
6.1	Sequential Diagnosis of A New Patient	128

CHAPTER 1

INTRODUCTION

Today's hospitals encompass a broad range of computer applications, from administrative data processing to automated medical methods. In the latter are such applications as on-line monitoring of patients in intensive care units and automatic selection of most-compatible organ recipients.

Particularly the computer is being applied to automatic interpretation of ECGs and, through automated medical interviews, to mass screening. From here it is only one more step to automated medical diagnosis. That step has yet to be taken, at least so far as implementation is concerned. Current research is attempting to determine a method for automatic medical diagnosis which will lead to consistently satisfactory results. It is this area of research which is examined in this thesis.

When a doctor examines a patient and attempts to determine which disease the patient has, the doctor first obtains items of information from the patient's history, physical examination and laboratory tests (Ledley, 1959). Secondly he assesses these items of information in the light of his knowledge of the group of diseases from which he thinks the patient may be

suffering (Taylor, 1970). This process is often sequential since the doctor revises his opinion about the diagnosis as new items of information become available. Finally either one disease is diagnosed, or treatment is commenced without a definitive diagnosis (Wang, 1972).

The items of a patient's history, physical examination and laboratory tests all produce single items of information called symptoms, signs and tests respectively. However, for the purpose of this thesis the terms symptoms, signs and tests are considered synonymous. Thus the diagnostic process involves knowledge of a large volume of diseases, and the relationship that exists between symptoms and disease. It also involves the matching of the patient's symptoms with the symptoms of all the possible diseases.

It is generally considered that it is the omission of data from this process that most frequently leads to errors in diagnosis. The most brilliant physician is always the one who remembers and considers the most possibilities (Clendening, 1947).

With the above realizations it is inevitable that the large data-handling capabilities of the computer have been applied to medical diagnosis. A review of some of these applications is presented in Chapter 2.

In Chapter 3 a method is developed by which the computer can be used to determine the complex relationships which exist between symptoms and disease. These relationships, called here disease-symptom functions, are determined in such a manner as to overcome many of the assumptions of other methods.

Although disease-symptom functions can be used to diagnose any patient, it is preferred to consider with what probability each patient has certain diseases. Suitable methods for determining these probabilities are developed in Chapter 4.

In Chapter 5 the results, obtained using the methods developed in Chapters 3 and 4, are compared with results obtained using other methods for automatic diagnosis. Reasons are given as to why superior results are obtained using the methods developed herein.

If the current diagnosis is indefinite it may be possible to use additional symptoms. A methodology for sequential diagnosis is developed in Chapter 6.

A summary, conclusions and recommendations for further research are presented in Chapter 7.

CHAPTER 2

LITERATURE SURVEY

2.1 Introduction

This chapter reviews current methods for automatic medical diagnosis in preparation for the alternative methods developed in this thesis. Section 2.2 formally defines the mathematical notation to be used for patients, symptoms and disease. Section 2.3 then reviews Bayesian methods of diagnosis, a cardinal assumption of which is that, for each disease, the considered symptoms are independent. Corrections for this assumption are reviewed in section 2.4. Non-Bayesian methods are reviewed in section 2.5. Methods which allow for all orders of dependence between the symptoms of a disease are reviewed in section 2.6. Conclusions are presented in section 2.7.

Prior to the review, attention must be drawn to two facts. First all methods for automatic diagnosis use a data base which contains items of information regarding a set of previously diagnosed patients. This data base is primarily used to determine the relationships which exist between symptoms and disease. Unfortunately the field of automatic medical diagnosis is restricted by a lack of suitable data bases.

In this connection Croft (1972) has suggested the formation of a liaison group to establish large reliable medical data bases. In the meantime, researchers must gather their own statistics. This is not a very straightforward matter, for medical records are private and not readily accessible. In consequence, most researchers are restricted to working in conjunction with an individual hospital researcher or department, resulting in data being gathered relative to only one category of disease. This limits the extent to which automatic methods can be tested, since different diseases exhibit different complexities.

The second fact concerns the terminology used in this thesis when referring to patients. Patients used to generate the data base will be called "previous patients"; their symptoms and associated disease are known. In contrast a "new patient" will be one who exhibits a combination of symptoms which is different to that of every previous patient. The word "patient" will be used to refer to someone whose symptoms may, or may not, be equal to those of any patient in the data base. Clearly new patients are the most difficult to diagnose. Unfortunately few researchers have attempted to diagnose new patients.

2.2 Patients, Symptoms, Diseases

Let N be the total number of all previously diagnosed patients. Then each previous patient can be identified by a number n, where n is a member of the set

$$\Pi = \{1,2,....,n,....,N\} \quad . \tag{2.1}$$

Let the previous diagnoses depend on the observation of the set of M symptoms

$$S = \{S_1,S_2,....,S_m,....,S_M\} \quad . \tag{2.2}$$

Then the association between each previous patient n and the symptoms can be described by a symptom vector

$$S(n) = \{S_1(n),...,S_m(n),....,S_M(n)\} \tag{2.3}$$

where the magnitude of $S_m(n)$ is a measure of the extent to which the m'th symptom is observed in the n'th previous patient.

For a given value of n, the symptom vector S(n) corresponds to what Ledley has termed the "symptom complex" of the n'th patient. Rinaldo (1963) has regarded S(n) as constituting a "symptom profile" of the n'th patient.

Suppose further that from these observations each previous patient has been clinically diagnosed as having one disease $D(n) = D_k$. Then the set of all such

diseases is

$$D = \{D_1, D_2, \ldots, D_k, \ldots, D_K\} \tag{2.4}$$

where $K \leq N$. Then

$$D(n) \in D \quad \text{if} \quad n \in \Pi \ .$$

Hence every previous patient has a record

$$P_n = \{n;\ S(n);\ D(n)\} \tag{2.5}$$

with the set

$$P = \{P_1, P_2, \ldots, P_n, \ldots, P_N\} \tag{2.6}$$

forming a date base.

Note that the set Π may be partitioned into two disease sets Π^{k1} and Π^{k2}. The former is the set of all previous patients having the disease D_k, and the latter is the set of all previous patients not having the disease D_k. Therefore

$$\Pi = \Pi^{k1} + \Pi^{k2} \tag{2.7}$$

where

$$\Pi^{k1} = \{n | D(n) = D_k\} \tag{2.8}$$

and

$$\Pi^{k2} = \{n | D(n) \neq D_k\} \ . \tag{2.9}$$

A previous patient diagnosed as having the disease D_k will be denoted by n_{k1}. Thus

$$n_{k1} \in \Pi^{k1} \tag{2.10}$$

and

$$D(n_{k1}) = D_k \quad .$$

A previous patient, diagnosed as not having the disease D_k, will be denoted by n_{k2}. Thus

$$n_{k2} \in \Pi^{k2} \tag{2.11}$$

and

$$D(n_{k2}) \neq D_k \quad .$$

The number of elements in set Π^{k1} will be denoted by N_{k1}; similarly the number of elements in the set Π^{k2} will be denoted by N_{k2}. Therefore

$$N = N_{k1} + N_{k2} \quad . \tag{2.12}$$

A new patient will be denoted by n^*.

2.3 Bayesian Methods of Diagnosis

The data base can be used to estimate the probability $P(S(n)|D_k)$, and what is required for diagnosis is the probability $P(D_k|S(n))$. One solution to the problem of determining the latter is Bayes' theorem which follows from the multiplication rule of probability,

$$P(S(n).D_k) = P(S(n)).P(D_k|S(n)) \ .$$

Since

$$P(S(n).D_k) = P(D_k.S(n))$$

it follows that

$$P(S(n)).P(D_k|S(n)) = P(D_k).P(S(n)|D_k) \ ,$$

thus

$$P(D_k|S(n)) = \frac{P(D_k).P(S(n)|D_k)}{P(S(n))} \ . \qquad (2.13)$$

Since data is gathered in the form {n;S(n);D(n)} it is convenient to assume that the diseases are mutually exclusive. Then from the addition rule of probability

$$\sum_{k=1}^{K} P(D_k|S(n)) = \sum_{k=1}^{K} \frac{P(D_k).P(S(n)|D_k)}{P(S(n))} = 1.0$$

and (2.13) may be expressed in the form

$$P(D_k|S(n)) = \frac{P(D_k).P(S(n)|D_k)}{\sum_{k=1}^{K} P(D_k).P(S(n)|D_k)} \qquad (2.14)$$

which is the more frequently used version of Bayes' theorem[1].

---

[1] When the assumption of symptom independence is made (2.13) is no longer exact. Accordingly (2.13) is not a correct probability and numbers such as $P(D_k|S(n))=60.1$ are obtained. Normalization as in (2.14) is necessary to ensure that $\sum_{k=1}^{K} P(D_k|S(n)) = 1.0$.

Consider the terms in the numerator of (2.14) (the denominator simply serves as a normalization factor). The term $P(D_k)$ is the prior probability that the patient has the disease D_k, irrespective of the symptoms. It is this term which takes into account geographical location, seasonal epidemics, etc.

The second term in the numerator is

$$P(S(n)|D_k) = P(S_1(n),S_2(n).....S_M(n)|D_k)$$

$$= P(S_1(n)|D_k).P(S_2(n)|S_1(n).D_k)..... \quad (2.15)$$

$$....P(S_M(n)|S_1(n).S_2(n),....,$$

$$S_{M-1}(n).D_k) \ .$$

Since the conditionalities on the right of (2.15) require knowledge of the symptom combination as shown on the left this equality cannot be used for new patients. However, if, for each disease, the symptoms are assumed to be independent then (2.15) becomes

$$P(S(n)|D_k) = P(S_1(n)|D_k).P(S_2(n)|D_k)....P(S_M(n)|D_k) \ . \quad (2.16)$$

Since each of the right hand side probabilities can be estimated from the data base, (2.16) may be used when diagnosing new patients.

Four applications of Bayes' theorem, assuming symptom independence, will now be reviewed. The order

of presentation is logical rather than chronological.

2.3.1 The Work of Reale

Reale (1968) attempted automatic medical diagnosis using Bayes' theorem. With a data base of 1148 previous patients having a total of 94 different congenital heart diseases and exhibiting 46 different symptoms, Reale calculated the prior probabilities $P(D_k)$ and the conditional probabilities $P(S_m(n)|D_k)$.

Assuming that for each disease the symptoms were independent, each of these previous patients was diagnosed by listing the disease probabilities $P(D_k|S(n))$ in a decreasing order of magnitude. The computer diagnosis and the provisional clinical diagnosis, using the same symptoms S(n), were then compared with the final diagnosis. Coincidence with the correct diagnosis occurred in 73% of the cases with the clinical approach and in 81% with the Bayesian method. In short the computer had beaten the doctor by a margin of 8%.

The same method was applied to another group of patients whose symptoms and associated diseases had not been used to build up the data base. The computer accuracy then dropped to 60%, compared with 73% for the doctor. Reale felt that this drop in accuracy could be explained by the different prior probabilities $P(D_k)$ for the other group of patients.

2.3.2 The Work of Warner

Warner (1961) was one of the early researchers in the field of automatic medical diagnosis. He worked with a data base of 1,035 previous patients that included 33 different congenital heart diseases.

Warner took great care to see that Bayes' theorem was used correctly. Of 50 symptoms observed, only 31 symptoms were considered sufficiently independent for use in the theorem. The symptoms were obtained from findings in X-rays, ECGs, heart murmurs, and phonocardiographic tracings. Also, Warner frequently observed that the absence of a symptom is as significant as a presence. Warner investigated a "modified" version of Bayes' theorem in which if the symptom $S_m(n)$ is absent the term $P(S_m(n)|D_k)$ is not used in the formulation.

Warner did not provide figures for the percentage accuracy of his results. He maintained that the 36 additional patients whom he diagnosed were too few for a full evaluation. However, he found that the correct version of Bayes' theorem gave more accurate results than the modified version.

Further Warner went on to show how the exclusion of certain symptoms can significantly alter the diagnosis, both for the better and for the worse. He argued that the selection of symptoms for consideration in studies of this sort must be done with extreme care.

2.3.3 The Work of Boyle

When Reale attempted to diagnose patients who were not part of the data base, he found that the diagnostic accuracy dropped considerably. Boyle (1966) attempted to overcome this problem.

The diagnosis of 300 consecutive cases of goitre were used to determine the prior probabilities of the diseases. These were Hashimoto's disease 0.1, simple goitre 0.89, and thyroid cancer 0.01. The conditional probabilities of the symptoms, under the assumption of their independence, were obtained from 51 previous patients with simple goitre, 53 previous patients with Hashimoto's disease, and 51 previous patients with thyroid cancer.

A further 88 patients were used to compare clinical with automatic diagnoses. Both diagnoses were based upon 30 different observations of clinical signs and laboratory tests performed on all patients.

The provisional clinical diagnostic accuracy was 77%. With the prior probability terms included Bayes' theorem gave an accuracy of 83%. But without the prior probability terms this figure increased to 85%.

Boyle argued that the increase in diagnostic accuracy, obtained by ignoring the prior probabilities,

is significant. He supported this argument by observing that the prior probabilities are highly dependent upon the population from which the patients are selected. Since the composition of these populations (say between a doctor's surgery and a special clinic at a hospital) varies drastically in terms of probability of disease, it is extremely difficult to calculate the prior probabilities accurately for any given patients. Even when all the patients are selected from one population, in this case a hospital clinic, Boyle maintained that the composition of this population will vary from day to day depending on which doctors are sending patients to that clinic.

Essentially Boyle showed that the prior probabilities are variable. In consequence, it may be better to assume them equal rather than to estimate them.

2.3.4 The Work of Scheinok

In most applications of automatic medical diagnosis all the symptoms are included when calculating the disease probabilities. Scheinok (1967) observed that it is often impractical, or at least costly, to determine whether or not a patient has all the possible symptoms. Some symptoms may be redundant and the cost

and inconvenience of determining their existence is wasted.

Scheinok decided to determine a subset of symptoms which would lead to as accurate a diagnosis as a full set. Additionally he incorporated a correction for small samples, as proposed by Bailey (1965). The need for the correction arises because, in terms of the binary-valued symptoms used in this application, some symptoms are consistently present, $S_i(n_{kl}) = 1$, others are consistently absent, $S_j(n_{kl}) = 0$. In such instances, Bayes' theorem diagnoses any new patient for whom $S_i(n^*) = 0$ or $S_j(n^*) = 1$ as not having the disease D_k. Bailey's correction prevents such absolute diagnoses from occurring.

The data gathered for the analysis was from 300 previous patients having a total of 6 different diseases and exhibiting 11 different symptoms, all relating to upper abdominal pain. Bayes' theorem was used assuming independence of the symptoms.

Scheinok worked by trial and error calculating the disease probabilities of the 300 previous patients using every combination of subset size from 3 to 11 symptoms. For each subset the combination yielding the highest diagnostic accuracy was selected as being the ultimate for that subset.

In this manner Scheinok determined that the diagnostic accuracy increased up to a subset size of 9 symptoms but not beyond that. The accuracy of diagnosis was then found to be 76.7%.

2.3.5 Comment

Bayes' theorem is by far the most popular method used in automatic diagnosis and the number of applications is high. For the purpose of this thesis the review has been limited to those researchers who have presented original ideas in its application. For a more detailed review the reader is referred to Wang (1972).

2.4 Corrections to Bayes' Theorem

The fact that Bayes' theorem does not give 100% accuracy when diagnosing previous patients shows that one, or more, of three assumptions is false. These assumptions are; the diseases are mutually exclusive; the symptoms are independent; each symptom vector is unique to one disease. This section reviews two attempts, made by Scheinok (1969), to correct for the assumption of symptom independence.

2.4.1 Vanderplas' Correction

Vanderplas (1967) has suggested that when two (or more) symptoms $S_i(n)$ and $S_j(n)$ are dependent then the conditional probabilities be determined from the relation

$$P(S_1(n),\ldots,S_i(n),S_j(n),\ldots S_M(n)\,|\,D_k)$$

$$= P(S_1(n)\,|\,D_k)\ldots.P(S_i(n)S_j(n)\,|\,D_k)\ldots P(S_M(n)\,|\,D_k).$$

The probability $P(S_i(n)S_j(n)\ D_k)$ is determined from a count of those previous patients who exhibit the symptom combination $S_i(n),S_j(n)$.

Note that the correction can only be used to diagnose new patients if the new patients exhibit the same symptom combination $S_i(n^*),S_j(n^*)$. Otherwise the symptoms must be assumed to be independent as before.

Scheinok applied the correction to his original group of 300 previous patients having a total of 6 different diseases and exhibiting 11 different symptoms relating to upper abdominal pain. Pairs of symptoms having statistically significant correlation coefficients were assumed to be dependent. All other symptoms were assumed to be independent.

The method produced no improvement in diagnostic accuracy compared with the uncorrected version of Bayes' theorem. Each gave 76.7% accuracy, although diagnosing

a different subset of previous patients as having each disease.

2.4.2 Best's Correction

Best (-) suggested that for the case of dependence the relation

$$P(S_1(n).....S_M(n)\,|D_k) = P(S_1(n)\,|D_k)....P(S_M(n)\,|D_k)$$

can be used, provided that each conditional probability on the right hand side of the equality is weighted by an exponent. He defined each exponent as being a function of the multiple correlation coefficient which relates that symptom with all the other symptoms as exhibited by that disease. Scheinok (1969) applied the method to the same data as before. He found that the method gave a 1% improvement in the accuracy of diagnosis of previous patients, compared with the uncorrected version of Bayes' theorem.

2.4.3 Discussion

Without being able to examine the results in detail it is difficult to determine why, in particular, Vanderplas' correction did not improve the results. However, of the 300 previous patients diagnosed 178 exhibited symptom vectors which were duplicated in two

or more diseases. Consequently Vanderplas' correction may have changed the diagnosis of the duplicates from one disease to another, thereby incorrectly diagnosing the duplicates elsewhere. The result could be no improvement in diagnostic accuracy.

With regard to Best's correction Scheinok observed that there are many methods of applying weights, besides exponential ones. Perhaps other methods would improve the results.

2.5 Non-Bayesian Methods of Diagnosis

While Bayes' therorem is the most popular method, in the field of automatic medical diagnosis, it has clearly been far from successful. Various other methods have been tried, and several are reviewed here.

2.5.1 Boolean Algebra

Ledley (1959, 1960) has formulated, though never applied, several non-Bayesian methods of automatic diagnosis. His methods use Boolean Algebra.

If it can definitely be stated that any patient having the disease D_k has the symptom S_m (or a certain combination of symptoms) then Boolean Algebra could be used with certainty. Unfortunately this is rarely the case for, in each disease set, each symptom is likely

to be present in only a certain proportion of the previous patients. Further the method is limited to only diagnosing, as having the disease D_k, those new patients who exhibit the same symptom (or combination of symptoms), as some previous patient with D_k.

2.5.2 Weighted Symptom Summation

Crooks (1959) has determined a "clinical diagnostic index" which distinguishes between non-toxic and thyrotoxic patients. The index is obtained by summation of a set of weights according to the presence and absence of each set of symptoms.

Crooks used 23 symptoms relating to the clinical diagnosis of thyrotoxicosis. By applying the method to 99 non-toxic and 83 thyrotoxic previous patients, suitable weights were determined to obtain statistically significant separation between the indices for the two types of previous patients. The method was then applied to another group of 121 patients and achieved 85% diagnostic accuracy.

2.5.3 Discriminant Analysis

Scheinok (1968) diagnosed his original group of 300 previous patients using a method known as discriminant analysis. The method uses the value of the

weighted sum of the symptoms to classify each patient. Whereas Crooks determined suitable weights by trial and error, Scheinok used an algorithmic method developed by Fisher (1936). A different set of weights is determined for each disease, and the patient is diagnosed as having that disease for which the sum of the weighted symptoms is the largest.

One advantage of the method is that the symptoms can be multivalued. This is often important since disease is a dynamic process and what appears to be a minor symptom may be indicative of a disease not yet fully developed.

Scheinok again determined the subset of symptoms which would yield the highest diagnostic accuracy. He found that the entire set of 11 symptoms produced the highest accuracy. This accuracy was 75% compared with 76.7% when using Bayes' theorem.

2.5.4 Least-Squares-Fit

In a method proposed by Heaps (1973) each disease has its own disease-symptom function, being any mathematical expression of the symptoms. When diagnosing a patient the value of the disease-symptom function for each possible disease is determined from the symptoms which the patient exhibits. The diagnosis is made

according to which disease-symptom function gives the value closest to unity. A least-squares-fit method is used in an attempt to make the value of each disease-symptom function be unity for previous patients having that disease and zero for previous patients not having that disease.

The method was applied by the author (Cumberbatch, 1973) to data supplied by Scheinok. Using linear disease-symptom functions the results were only marginally inferior to those obtained by Scheinok using Bayes' theorem (76.2% compared with 76.7%).

Unfortunately the accuracy of the method was found to be dependent upon the scales used to quantize the symptoms.

2.5.5 K-Nearest-Neighbours Rule

A K-nearest-neighbours rule has been applied by Croft (1972) to the diagnosis of patients suffering from 20 different types of liver disease. The method determines the Euclidean distances between each patient to be diagnosed and all previous patients. These distances are used to find the K neighbours nearest to the patient to be diagnosed. The patient is then diagnosed as having that disease which is present in the largest number of these K neighbours.

A total of 1991 previous patients exhibiting 50 different multivalued symptoms were used to diagnose a further 437 patients. The accuracy of diagnosis varied with the number K as follows; 51% (K = 1), 62% (K = 10), 59% (K = 25). No value of K gave results superior to those of Bayes' theorem (64%), assuming symptom independence.

2.5.6 Discussion

The advantage of these non-Bayesian methods for automatic diagnosis is that they make no assumptions of symptom independence. Since this assumption is the prime cause of error when diagnosing previous patients it is intuitive to expect that non-Bayesian methods will give a higher accuracy of diagnosis of previous patients.

However,it can be shown, (see Chapter 5, also Duda and Hart (1973)) that all methods divide up the disease-symptom space into regions, one or more regions for each disease. A patient is diagnosed according to the region into which his symptom vector places him.

The methods differ in the criteria used to determine the separating surfaces which define these regions. For instance Crooks (1959), Scheinok (1968) and Cumberbatch (1973), by using linear functions of the symptoms, divided up the disease-symptom space with hyperplanes. Croft (1972), by using the K-nearest

neighbours rule, used non-linear separating surfaces, each region being redetermined for each patient to be diagnosed.

In all instances the resulting accuracy of diagnosis is dependent upon the orientation chosen for the separating surfaces. Further, the separating surface used must be suitable for the data.

2.6 Allowance for Dependence Between Symptoms

This section reviews two methods of automatic diagnosis in which allowance is made for quadratic, cubic etc. orders of dependence between the symptoms of a disease.

2.6.1 A Bayesian Method

Bahadur (1961) proposed a distribution which allows for the dependence of all orders between the symptoms of a disease. The method, however, involves massive calculations since correlations between all orders of symptoms must be determined for each disease. Scheinok (1972a) used the distribution in conjunction with Bayes' theorem.

Applying the method to the same data as before, Scheinok produced a lexicon of symptom vectors showing

the calculated probabilities for each disease. These probabilities simply equalled the frequency of occurrence of the symptom vectors within the data base. Thus the method had correctly diagnosed all previous patients; i.e. 100% diagnostic accuracy. Note that the lexicon also contained disease probabilities for symptom vectors not in the data base and it was not determined how accurate these were (e.g. by diagnosing new patients).

2.6.2 A Non-Bayesian Method

The disease-symptom functions proposed by Heaps may be expressed in terms of any combination of symptoms. The author (Cumberbatch, 1973) applied Heaps' method to the data supplied by Scheinok using quadratic disease-symptom functions.

The result was an overall diagnostic accuracy of 88.8%. Further, by prefiltering the data, this figure was raised to 92.4%.

2.6.3 Discussion

The high diagnostic accuracy obtained with both methods is not surprising. It is a property of both methods that as the order of dependency between the symptoms of a disease is increased (quadratic, cubic, etc.) so is the resulting accuracy of diagnosing

previous patients.

The reason for the increase in accuracy is that a non-linear model uses non-linear surfaces to divide up the disease-symptom space (see Chapter 5). If the order of non-linearity is suitably increased and the resulting separating surfaces are suitably oriented then the accuracy of diagnosis can be made to approach 100%. Indeed Davis (1972) has mathematically proven that if all orders of dependence between the symptoms of a disease are used then, in particular, Bahadur's distribution in conjunction with Bayes' theorem leads to 100% diagnostic accuracy.

2.7 Conclusions

When comparing different methods of diagnosis it is important to consider the number and type of symptoms used. Warner, for instance, used 31 different symptoms obtained from findings in X-rays, ECGs, heart murmurs, etc. Scheinok, however, used only 11 symptoms and these were obtained by asking the patients questions and recording their yes-no type answers. Clearly it is difficult to compare these results unless each can be compared with the accuracy of diagnosis obtained by a doctor when using the same symptoms. Only Reale and Boyle provided such information.

The accuracy of diagnosis of previous patients is not a suitable measure by which methods

can be compared. By making suitable corrections for the assumption of symptom independence Bayes' theorem will diagnose previous patients with increasingly higher accuracy (Vanderplas, Best). Both Bayesian and non-Bayesian methods will diagnose previous patients more accurately as the order of dependence between the symptoms is increased (Bahadur, Heaps).

But there is no need to apply automatic methods of diagnosis to previous patients. For the diagnosis of such patients may be made directly from the frequency of occurrence of each previous patient's symptom vector within the data base.

When using automatic methods of diagnosis it must be remembered that the relationship between symptoms and disease as exhibited by the previous patients may not be the same as that exhibited by new patients. Thus the accuracy of diagnosis of previous patients is only indicative of that which might be obtained when diagnosing new patients.

The foregoing review has outlined attempts to formulate and apply methods for automatic diagnosis. It is clear that the accuracy of diagnosis is dependent upon many factors and that there is still considerable scope for research. The following chapters relate to

the formulation and application of alternate methods for automatic diagnosis.

CHAPTER 3

DISEASE-SYMPTOM FUNCTIONS

3.1 Introduction

The methods for automatic diagnosis reviewed in sections 2.5.2, 2.5.3 and 2.5.4 all use the value of the weighted sum of the patient's symptoms for diagnosis. Different weights are used for each disease and by determining weights for symptom pairs, symptom triplets, etc., diseases may be assumed to depend non-linearly upon the symptoms. Heaps (1973) regarded the resulting relation as being the "k'th disease-symptom function".

The particular advantage of all such methods is that they make no assumptions as to the independence of the symptoms. Further, once the weights have been determined, the diagnosis of any patient is relatively straightforward. Indeed Freeman (1972) has determined that some physicians use weighted-symptom summation when making clinical diagnoses.

These methods, however, share a disadvantage with those based on Bayes' theorem. No methods allow for the lowering of the certainty of some correct diagnoses in an attempt to increase the certainty of others.

In the approach presented here (also Cumberbatch, 1974), the k'th disease-symptom function is chosen to assume a maximum value for all previous patients having the k'th disease, and a minimum value for all previous patients not having the k'th disease. The scale of these values may be linearly changed by application of a parameter α_k. This permits a better distinction to be made between several functions which assume a maximum value for previous patients known to have only one disease.

A second parameter, β_k, may be used to force some degree of consistency on the values of the k'th disease-symptom function for previous patients having the k'th disease. This is achieved by changing the ratio of the standard deviation to the mean of these values.

The disease-symptom functions take into account both the presence and the absence of the symptoms, a fact which Warner noted as providing useful information. Yet they are not dependent upon the quantization of the symptoms, which was the case with Heaps' method. Also, multivalued symptoms may be used, which Scheinok (1968) observed as often useful. However, they are restricted to diagnosing any patient as having only one disease.

The method has been applied to data supplied by Scheinok. Results are presented in section 3.8. The

accuracy of diagnosis of previous patients is shown to be as high as 80.3%. New symptom vectors are diagnosed using linear and quadratic disease-symptom functions. The resulting accuracy of diagnosis is examined in relation to the growth of the data base.

3.2 Disease-Symptom Functions

Any relationship may be assumed to exist between symptoms and the disease D_k. Particularly if this relationship is assumed to be linear then the k'th disease-symptom function is given by

$$Z_k(n) = \sum_{m=1}^{M} C_{km} S_m(n) + C_{ko} \qquad (3.1)$$

where the C_{km} and C_{ko} are the coefficients (or weights) of the linear process.

In the (M+1)-space (i.e. M symptoms and $Z_k(n)$) (3.1) represents a linear surface (hyperplane) and the symptom vectors S(n) represent points in this space. If the coefficients in (3.1) are suitably chosen then the intersection of the resulting linear surface with the symptom space might separate all points in the disease set $n \in \Pi^{k1}$, $D(n) = D_k$ from points in the disease set $n \in \Pi^{k2}$, $D(n) \neq D_k$, as illustrated in Figure 3.1.

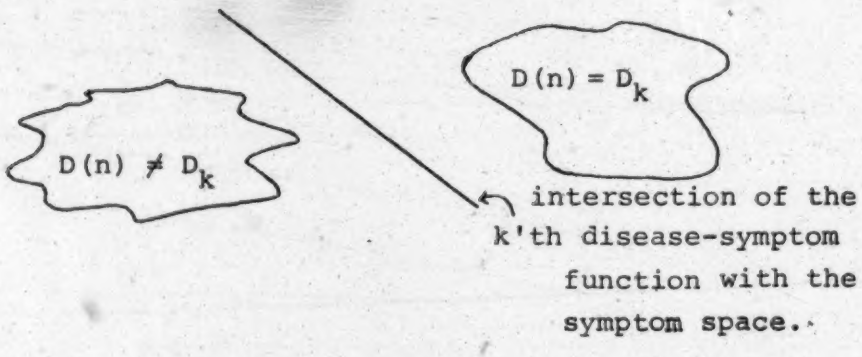

Figure 3.1. Linear Separation of Two Disease Sets.

In order to discriminate between K disease sets, K disease-symptom functions must be used. Then any point p in the entire symptom space may be classified (see Figure 3.2) according to the decision rule

$$\text{if } Z_k(p) > Z_i(p), \quad \text{all } i \neq k \qquad (3.2)$$

$$\text{then } D(p) = D_k .$$

(In the case of ties, either the classification is undefined or the patient has more than one disease.)

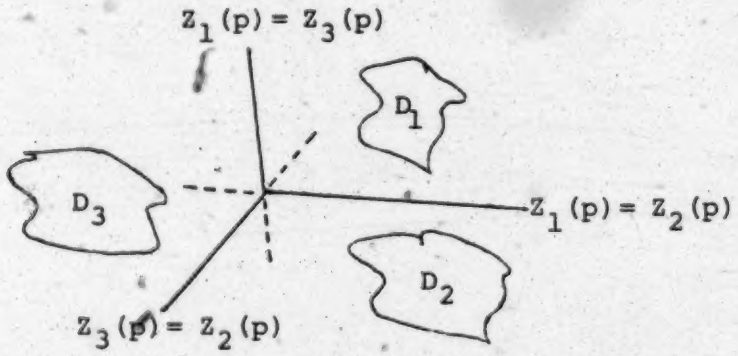

Figure 3.2. Linear Separating Surfaces for Three Disease Sets.

Note that by increasing the order of dependence between the symptoms and the disease non-linear separating surfaces may be used.

Several authors (Fisher, 1936; Sebestyen, 1962; Nilsson, 1965; Heaps, 1973) have formulated methods for determining coefficients for use in (3.1). Particularly Carl and Hall (1972) have suggested that the k'th disease-symptom function be regarded as a filter designed to separate the signals

$$z^{k1} = \{z_k(n);\ n \in \Pi^{k1}\}$$

from the noise

$$z^{k2} = \{z_k(n);\ n \in \Pi^{k2}\} .$$

They proposed determining the coefficients of (3.1) using the Wiener filter technique (see Levinson (1947)).

From the point of view of linearly changing the scale of the $z_k(n)$ a more suitable filter is that proposed by Dwork (1950). With one modification[1] the method maximizes the expression

$$R_k = \frac{\frac{1}{N_{k1}} \sum_{n_{k1}} z_k(n_{k1})}{[\frac{1}{N} \sum_{n} z_k^2(n)]^{\frac{1}{2}}} \qquad (3.3)$$

being the ratio of the average amplitude of the signals to the root mean square of the signals plus noise.

---

(1) Dwork originally proposed maximizing the ratio of the average amplitude of the signals, to the root mean square of the noise. The modification (3.3) imposes the additional condition (iii) of (3.9), which forces consistency on the amplitude of the signals.

Let

$$\bar{Z}_{k1} = \frac{1}{N_{k1}} \sum_{n_{k1}} Z_k(n_{k1}) \tag{3.4}$$

and

$$\bar{Z}_{k2} = \frac{1}{N_{k2}} \sum_{n_{k2}} Z_k(n_{k2}) \quad . \tag{3.5}$$

Then since

$$\sum_{n_{k1}} Z_k^2(n_{k1}) = \sum_{n_{k1}} (Z_k(n_{k1}) - \bar{Z}_{k1})^2 + N_{k1}\bar{Z}_{k1}^2 \tag{3.6}$$

and

$$\sum_{n_{k2}} Z_k^2(n_{k2}) = \sum_{n_{k2}} (Z_k(n_{k2}) - \bar{Z}_{k2})^2 + N_{k2}\bar{Z}_{k2}^2 \tag{3.7}$$

substitution into (3.3) shows that

$$R_k = \frac{\bar{Z}_{k1}}{[\frac{1}{N}(\sum_{n_{k1}} Z_k^2(n_{k1}) + \sum_{n_{k2}} Z_k^2(n_{k2})]^{\frac{1}{2}}}$$

$$= \left(\frac{\sum_{n_{k1}} (Z_k(n_{k1}) - \bar{Z}_{k1})^2}{N\,\bar{Z}_{k1}^2} + \frac{N_{k1}}{N} + \frac{\sum_{n_{k2}} (Z_k(n_{k2}) - \bar{Z}_{k2})^2}{N\,\bar{Z}_{k1}^2} + \frac{N_{k2}\bar{Z}_{k2}^2}{N\,\bar{Z}_{k1}^2} \right)^{-\frac{1}{2}} \tag{3.8}$$

Inspection of (3.8) shows that R_k achieves its true maximum value of $\sqrt{N/N_{k1}}$ if the $Z_k(n)$ are chosen so that

$$\left.\begin{array}{lll} \text{(i)} & \bar{Z}_{k1} > 0 & \text{(by (3.3) and (3.8))} \\ \text{(ii)} & \bar{Z}_{k2} = 0 & \text{(by (3.8))} \\ \text{(iii)} & Z_k(n_{k1}) = \bar{Z}_{k1}, \text{ all } n_{k1} & \text{(by (3.8))} \\ \text{(iv)} & Z_k(n_{k2}) = \bar{Z}_{k2}, \text{ all } n_{k2} & \text{(by (3.8))} \end{array}\right\} \tag{3.9}$$

If conditions (3.9) are satisfied for all $k \leq K$ then the decision rule (3.2) will correctly classify all previous patients. It is therefore appropriate to choose each k'th disease-symptom function in such a manner as to maximize R_k.

The disease-symptom functions can assume any order of dependence between the symptoms and the disease. Let this dependence be denoted by $f_k(S(n))$. Then condition (i) of (3.9) shows that there is no upper bound on the value of each $\bar{Z}_{k1}$. Accordingly maximization of R_k leads to solutions for the $f_k(S(n))$ of the form

$$Z_k(n) = \alpha_k f_k(S(n)) \tag{3.10}$$

where α_k is a scalar quantity.

In application, the form of the disease-symptom functions $f_k(S(n))$ prevents the true maximum value of R_k from being obtained. However, conditions (3.9) imply that there may exist some Z_k^*, such that

$$\begin{aligned} &Z_k^* \leq Z_k(n_{k1}) \quad , \quad \text{all } n_{k1} \\ &\text{and} \\ &Z_k^* > Z_k(n_{k2}) \quad , \quad \text{all } n_{k2} \quad . \end{aligned} \tag{3.11}$$

If such a Z_k^* exists for all $k \leq K$ then, since (3.11) is independent of α_k, each α_k can be chosen so that

$$Z_1^* = Z_2^* = \ldots = Z_k^* = \ldots = Z_K^* \quad . \tag{3.12}$$

Then the decision rule (3.2) will correctly classify all previous patients.

If no such Z_k^* can be found, for all $k \leq K$, then the scalars α_k can be chosen so that decision rule (3.2) correctly classifies a maximum number of previous patients. In such instances the decision rule (3.2) may be satisfied by only a few of the previous patients. The reason, common to all previously formulated methods, is that the maximization of R_k can be greatly influenced by a few large values of some $Z_k(n_{k1})$. In such instances it is appropriate to increase the value of some $Z_k(n_{k1})$ at the expense of obtaining reduced values of other $Z_k(n_{k1})$. Such a transformation can be obtained by reducing the standard deviation of the $Z_k(n_{k1})$,

$$\sigma_{k1} = \left[\frac{1}{N_{k1}} \sum_{n_{k1}} (Z_k(n_{k1}) - \bar{Z}_{k1})^2\right]^{\frac{1}{2}} . \qquad (3.13)$$

However, σ_{k1} can be changed by the scalar α_k. Hence it is more appropriate to reduce the ratio of the standard deviation to the mean of the $Z_k(n_{k1})$,

$$r_k = \frac{\sigma_{k1}}{\bar{Z}_{k1}} = \frac{\left[\frac{1}{N_{k1}} \sum_{n_{k1}} (Z_k(n_{k1}) - \bar{Z}_{k1})^2\right]^{\frac{1}{2}}}{\frac{1}{N_{k1}} \sum_{n_{k1}} Z_k(n_{k1})} . \qquad (3.14)$$

Form (3.8) shows that r_k is related to R_k by the relation

$$R_k = \left[\frac{N_{k1} r_k^2}{N} + \frac{N_{k1}}{N} + \frac{\sum_{n_{k2}} Z_k^2(n_{k2})}{N \bar{Z}_{k1}} \right]^{-\frac{1}{2}} . \qquad (3.15)$$

If R_k is maximized subject to the constraint that r_k is a constant, (3.15) shows that changing the constant is analagous to changing the ratio N_{k1}/N used to weight r_k^2. Thus the maximization of R_k can be constrained, placing more (or less) emphasis on the term r_k^2 as is required.

This constrained maximization of R_k is most conveniently achieved by consideration of the expression

$$Q_k = R_k - \beta_k r_k . \qquad (3.16)$$

The constant β_k may be changed according to the emphasis placed on the minimization of r_k. The solution is then of the form

$$Z_k(n) = \alpha_k f_k(S(n), \beta_k) , \qquad (3.17)$$

where α_k and β_k are independent parameters. Accordingly the method may be referred to as the alpha-beta method of determining the coefficients of the disease-symptom functions.

Not only may the values of α_k and β_k be chosen differently for each disease, but the functions $f_k(S(n))$

may also be chosen differently. Such flexibility may be used so that the decision rule (3.2) correctly classifies a maximum number of previous patients. If the previous patients exhibit disease-symptom relations which are representative of the diseases, a maximum number of new patients will also be correctly diagnosed.

3.3 Linear Disease-Symptom Functions

Consider the special case in which the disease-symptom functions are assumed to be of the form

$$Z_k(n) = \sum_m C_{km} S_m(n) \quad . \tag{3.18}$$

Then (3.3) takes the form

$$R_k = \frac{\frac{1}{N_{k1}} \sum_{n_{k1}} \sum_m C_{km} S_m(n_{k1})}{[\frac{1}{N} \sum_n (\sum_m C_{km} S_m(n))^2]^{\frac{1}{2}}}$$

$$= \frac{\sum_m C_{km} \bar{S}_{km}}{[\sum_\ell \sum_m C_{k\ell} C_{km} \bar{S}_{\ell m}]^{\frac{1}{2}}} \tag{3.19}$$

where

$$\bar{S}_{km} = \frac{1}{N_{k1}} \sum_{n_{k1}} S_m(n_{k1}) \tag{3.20}$$

and

$$\bar{S}_{\ell m} = \bar{S}_{m\ell} = \frac{1}{N} \sum_n S_\ell(n) S_m(n) \quad . \tag{3.21}$$

Similarly (3.14) may be expressed in the form

$$r_k = \frac{\left[\frac{1}{N_{k1}} \sum_{n_{k1}} \left(\sum_m C_{km}(S_m(n_{k1}) - \bar{S}_{km})\right)^2\right]^{\frac{1}{2}}}{\sum_m C_{km}\bar{S}_{km}}$$

$$= \frac{\sum_\ell \sum_m g_{k\ell m} C_{k\ell} C_{km}}{\sum_m C_{km}\bar{S}_{km}} \qquad (3.22)$$

where

$$g_{k\ell m} = g_{km\ell} = \frac{1}{N_{k1}} \sum_{n_{k1}} S_\ell(n_{k1}) S_m(n_{k1}) - \bar{S}_{k\ell}\bar{S}_{km} \, . \qquad (3.23)$$

For given symptoms $S_m(n)$, all m, n, the expression

$$Q_k = R_k - \beta_k' r_k \qquad (3.24)$$

is stationary when

$$\frac{\partial Q_k}{\partial C_{km}} = \frac{\partial R_k}{\partial C_{km}} - \beta_k' \frac{\partial r_k}{\partial C_{km}} = 0 \, , \text{ all m.} \qquad (3.25)$$

It may be shown (see Appendix 1) that this condition may be expressed in the form

$$\sum_\ell \left(\bar{S}_{\ell m} + \frac{\beta_k'}{r_k R_k^3} g_{k\ell m}\right) C_{k\ell} = \alpha_k \bar{S}_{km} \qquad (3.26)$$

where α_k is independent of ℓ and m.

Equation (3.26) may be expressed in matrix form as

$$(\bar{S} + \beta_k G_k) C_k = \alpha_k \bar{S}_k \qquad (3.27)$$

where

$$\beta_k = \frac{\beta_k'}{r_k R_k^3} \quad ,$$

and

$$\bar{S} = [\bar{S}_{\ell m}]_{M\times M} \quad , \quad G_k = [g_{k\ell m}]_{M\times M} \quad ,$$

$$C_k = [C_{km}]_{M\times 1} \quad , \quad \bar{S}_k = [\bar{S}_{km}]_{M\times 1} \quad .$$

The solution is thus given by

$$C_k = \alpha_k (\bar{S} + \beta_k G_k)^{-1} \bar{S}_k \quad . \qquad (3.28)$$

For any α_k and β_k, equation (3.28) determines the coefficients C_{km}, for each linear disease symptom function (3.18), which produce a stationary value of Q_k.

R_k is the square root of a Rayleigh quotient, while r_k is the inverse of the square root of a Rayleigh quotient. The properties of Rayleigh quotients (Duda and Hart, 1973, p.117) lead to the conclusion that the solution (3.28) determines the coefficients C_{km}, for the linear disease-symptom function (3.18), which maximize Q_k.

The $\bar{S}_{km}$ and $g_{k\ell m}$ depend only on the symptoms of previous patients in the disease set Π^{kl}. However, $\bar{S}_{\ell m}$ measures the extent to which the symptoms S_ℓ and S_m occur in the set Π.

If the matrix $(\bar{S} + \beta_k G_k)$ is of reasonable size and well conditioned then the inversion of the matrix does not provide any computational difficulty. If the matrix is not well conditioned then one, or more, symptoms are probably dependent on the other symptoms. The problem is to determine which are the offending symptoms and to decide whether to remove them or not. This is a classical problem in matrix inversion and there are partial solutions.

3.4 Symptom Quantization

It is important that the $Z_k(n)$ be independent of the quantizing (scale of measure) of the symptoms. Hence each $Z_k(n)$ must be independent of any transformation of the form

$$S'_m = a_m S_m + b_m$$

for which

$$Z_k(n) = \sum_m a_m C_{km} S_m(n) + \sum_m b_m C_{km}$$

$$= \sum_m C'_{km} S_m(n) + C'_{ko} \quad . \qquad (3.29)$$

Accordingly if (3.18) is modified to include a constant, C_{ko}, then the resulting disease-symptom function

$$Z_k(n) = \sum_m C_{km} S_m(n) + C_{ko} \qquad (3.30)$$

is independent of the quantizing of the symptoms.

In the particular instance that the symptoms are binary valued, coefficients C_{kmo} and C_{kml} can always be found so that (3.31) can be written in the form

$$Z_k(n) = \sum_m C_{kml} S_m(n) + \sum_m C_{kmo}(1 - S_m(n)) \; . \qquad (3.31)$$

Thus the inclusion of the constant is seen to take into account the absence of symptoms, which has also been observed as useful for diagnosis.

The constant C_{ko}, is most easily included in (3.18) by addition of a redundant symptom, $S_o(n)$, common to all previous patients. All subsequent references to (3.18) will assume that the redundant symptom, $S_o(n)$, is included.

3.5 Generalized Linear Disease-Symptom Functions

A generalized linear disease-symptom function may be defined as one which is linear in the coefficients, yet not necessarily linear in the symptoms. Within this class of functions there are four categories (Wilson, 1973). By way of illustration consider the

four functions below, in which Z_k is dependent upon two symptoms S_1 and S_2.

(i) Linear and Non-Interactive Dependence

$$Z_k(n) = C_{k1}S_1(n) + C_{k2}S_2(n) + C_{ko}$$

(ii) Non-Linear and Non-Interactive Dependence

$$Z_k(n) = C_{k1}S_1(n) + C_{k11}S_1^2(n) + C_{k2}S_2(n) + C_{k22}S_2^2(n) + + C_{ko}$$

(iii) Linear and Interactive Dependence

$$Z_k(n) = C_{k1}S_1(n) + C_{k2}S_2(n) + C_{k12}S_1(n)S_2(n) + C_{ko}$$

(iv) Non-Linear and Interactive Dependence

$$Z_k(n) = C_{k1}S_1(n) + C_{k11}S_1^2(n) + C_{k2}S_2(n) + + C_{k22}S_2^2(n) + C_{k12}S_1(n)S_2(n) + C_{ko}.$$

The non-linear terms involve the square of the symptoms, and the interactive terms involve the product of the symptoms.

All these functions are linear in the coefficients. Accordingly the analysis of the preceding sections may still be applied. The only modification which must be made is an extension of the summation with respect to the index m in C_{km}, as appropriate.

The order of the non-linear and the interactive terms may be increased from quadratic to cubic etc.

Thus, for the purpose of this thesis, disease-symptom functions of category (i) will be said to be "linear", while those of categories (ii), (iii), and (iv) will be said to be "quadratic", "cubic", etc. as appropriate. Particularly, the general form of the quadratic disease-symptom function is given by

$$Z_k(n) = \sum_m C_{km} S_m(n) + \sum_{\ell \le m} \sum C_{k\ell m} S_\ell(n) S_m(n) \quad . \qquad (3.32)$$

In the instance that the symptoms are binary valued, the generalization

$$Z_k(n) = \sum_{m=1} C_{km} S_m(n) + \sum_{\ell \le m} \sum C_{k\ell m} S_\ell(n) S_m(n) + \ldots\ldots\ldots + C_{k1\ldots M} S_1(n) \ldots S_M(n) \qquad (3.33)$$

has the property that the coefficients can then be found such that $Z_k(n_{k1}) = 1$ and $Z_k(n_{k2}) = 0$ for any 2^M different symptom vectors of form (2.3). If the number of different symptom vectors is less than 2^M the inappropriate terms may be disregarded, and the coefficients again found such that $Z_k(n_{k1}) = 1$ and $Z_k(n_{k2}) = 0$. The decision rule (3.2) then correctly classifies all previous patients. This is consistent with Scheinok's (1972a) application of Bahadur's distribution to Bayes' theorem. However, as is shown in section 3.8.2, the correct diagnosis of all previous

patients does not necessarily lead to the correct diagnosis of all new patients.

3.6 A First Estimate for Alpha

With the parameters $\alpha_k = 1$ and $\beta_k = 0$ the coefficients of the k'th linear disease-symptom function (3.18) are given by the solution to the matrix equation

$$\bar{S}.C_k = \bar{S}_k \quad ,$$

by (3.28). Hence

$$\begin{aligned} \bar{Z}_{k1} &= C_k^T.\bar{S}_k \\ &= C_k^T.\bar{S}.C_k \\ &= \frac{1}{N}\sum_n Z_k^2(n) \quad , \end{aligned} \tag{3.34}$$

by (3.18) and (3.21). Therefore

$$R_k^2 = \frac{\bar{Z}_{k1}^2}{\frac{1}{N}\sum_n Z_k^2(n)} = \bar{Z}_{k1} \quad , \quad (\alpha_k=1,\ \beta_k=0) \tag{3.35}$$

Since R_k^2 has a maximum value of N/N_{k1} it follows that the maximum value of $\bar{Z}_{k1}$, $(\alpha_k = 1,\ \beta_k = 0)$, is also N/N_{k1}.

The decision rule (3.2) requires that the $Z_k(n)$, all $k \leq K$, be compared on a relative scale. Thus, as a first estimate,

$$\alpha_k = \frac{N_{k1}}{N} \quad . \tag{3.36}$$

Accordingly determination of the coefficients C_{km}, for the linear disease-symptom function (3.18), involves solving (3.28) with $\alpha_k = N_{k1}/N$ and $\beta_k = 0$. The parameters α_k and β_k are then changed, to provide an appropriate linear and non-linear scaling of the $Z_k(n)$, so that the decision rule (3.2) correctly classifies a maximum number of previous patients.

3.7 Confidence Limits

Suppose that the true, but unknown, accuracy rate of any method for automatic diagnosis of disease is b, and that when using this method k of N test samples are correctly diagnosed. Then k follows the binomial distribution and the fraction of test samples correctly diagnosed is exactly the estimate for b

$$\hat{b} = k/N \quad . \tag{3.37}$$

It is well known that $\hat{b}$ is a normal variable of mean b and standard deviation $\sqrt{b(1-b)/N}$. Hence if (3.37) is used to estimate the standard deviation of $\hat{b}$, the 95% confidence limits for b are

$$\hat{b} - 1.96\sqrt{\hat{b}(1-\hat{b})/N} \leq b \leq \hat{b} + 1.96\sqrt{\hat{b}(1-\hat{b})/N}. \tag{3.38}$$

When commenting on the relative accuracy rates of the different methods of diagnosis used in this thesis (3.38) will be used to determine whether one method is significantly better than another (at the 95% confidence level).

3.8 Results Using Disease-Symptom Functions

Data for this application was supplied by Scheinok(1972b). The data relates to 300 patients each suffering from one of six diseases: hiatal hernia duodenal ulcer, gastric ulcer, cancer, gallstones and functional disease. The physicians' determination of the first five diseases was through radiological examination of the stomach and gallbladder. The absence of any abnormality in the radiological studies was assumed to indicate the presence of functional disease. The result of the physicians' diagnosis is to allow a number to be assigned to the D(n) of (2.4) to indicate which of the six diseases each patient has.

The symptoms were based on the patients' answers to 11 questions chosen by physicians experienced in the diagnosis of the six diseases. If the n'th patient's reply to the n'th question was "yes", the $S_m(n)$ of (2.3) was set to "1". If the reply was "no", the $S_m(n)$ was set to "0". Descriptions of the 11 symptoms are listed by Scheinok under the categories of male, epigastric pain, right upper quadrant, back pain, clusters, brief irregular, food relief, food aggravation, positional aggravation, weight loss and persistence.

3.8.1 The Diagnosis of Previous Patients

The 11 binary valued symptoms, chosen by the

physicians experienced in the diagnosis of the six diseases, are not always sufficient to uniquely determine a patient's disease. Thus of the 300 previous patients there are only 122 who exhibit symptom vectors which are unique to one disease. The remaining 178 previous patients exhibit symptom vectors which are duplicated in two or more diseases.

Examination of the data reveals that when using decision rule (3.2) a maximum of 223 previous patients can be correctly diagnosed as having one of the six diseases. The remaining 77 patients will of necessity be incorrectly diagnosed. Thus, for the purpose of application of the alpha-beta method, these 77 previous patients were removed from the data base[(2)].

The first row of Table 3.1 shows the number of correct diagnoses of the 223 previous patients obtained by use of the least-squares-fit method, as proposed by Heaps (1973). The second row shows the corresponding numbers obtained by use of the alpha-beta method with $\alpha_k \neq 1$ and $\beta_k = 0$. The third row shows the corresponding numbers with $\alpha_k \neq 1$ and $\beta_k \neq 0$.

---

(2) The alternative approach would have been to define additional diseases, such that the symptom vectors of the 178 previous patients were unique to one disease.

	Hiatal Hernia	Duodenal Ulcer	Gastric Ulcer	Cancer	Gallstones	Functional Disease	Totals	Percentage Correct
Heaps(1973)	36	68	1	13	34	6	158	70.8
Alpha-Beta								
$\alpha_k \neq 1, \beta_k = 0$	36	68	9	11	33	16	173	77.6
$\alpha_k \neq 1, \beta_k \neq 0$	36	68	11	13	35	16	179	80.3
Number with disease	44	72	24	15	37	31	223	

Table 3.1 Number of Previous Patients Correctly Diagnosed When Using Linear Disease-Symptom Functions.

It may be noted that the least-squares-fit method is very poor in diagnosing previous patients with gastric ulcer and functional disease. The increase of 9.5 in the percentage accuracy of diagnosis, as obtained with the alpha-beta method, is achieved by improved diagnosis of previous patients with these diseases. There is no loss of accuracy of diagnosis of previous patients with other diseases.

When commenting on the accuracy of diagnosis of any automatic method it is also relevant to discuss the number of different symptom vectors correctly associated with their respective diseases. Any figure which involves only the total number of patients correctly diagnosed is not informative as to the flexibility of the method, for the correct diagnoses may have been obtained from only a few of the symptom vectors in each disease set Π^{kl}.

Accordingly Table 3.2 was prepared from the automatic diagnosis of the 223 previous patients to show the number of unique previous symptom vectors correctly associated with their respective diseases. In comparison with the least-squares-fit method the alpha-beta method gives an increase of 10.5 in the percentage accuracy.

	Hiatal Hernia	Duodenal Ulcer	Gastric Ulcer	Cancer	Gallstones	Functional Disease	Totals	Percentage Correct
Heaps (1973)	23	22	1	9	20	5	80	59.7
Alpha-Beta								
$\alpha_k \neq 1, \beta_k = 0$	23	23	6	8	20	9	89	66.4
$\alpha_k \neq 1, \beta_k \neq 0$	23	23	8	10	21	9	94	70.2
Number with disease	31	27	20	11	23	22	134	

Table 3.2 Number of Symptom Vectors Correctly Associated with Their Disease When Using Linear Disease-Symptom Functions.

It should be noted that if similar improvements in accuracy of diagnosis were obtained with the alpha-beta method, when diagnosing a random sample of 223 patients, the results would not be statistically significant at the 95% confidence level (see section 3.7).

3.8.2 The Diagnosis of New Symptom Vectors

For this application a total of 25 different symptom vectors were randomly removed from each of the six disease sets Π^{kl} and were regarded as new symptom vectors S(n*). New symptom vectors, rather than new patients, were used to prevent the accuracy of diagnosis from being influenced by a few multiply occurring symptom vectors.

The reduced data base was then similarly modified to contain 109 different symptom vectors S(n). This was done so that the accuracy of diagnosis of new symptom vectors could be examined in relation to the growth of the data-base size N.

The results of this application are shown in the graphs of Figures 3.3 and 3.4. Figure 3.3 shows the accuracy of diagnosis of previous symptom vectors using linear (points A, C, D) and quadratic (points B, E) disease-symptom functions, as the data-base size N

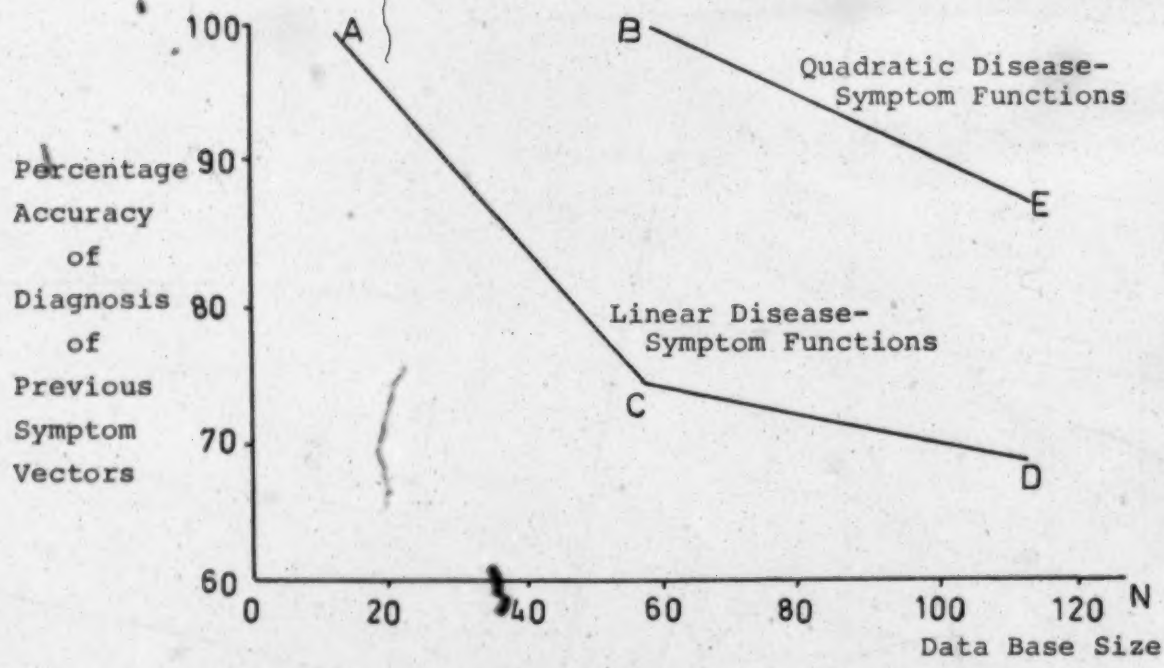

Figure 3.3 Percentage Accuracy of Diagnosis of Previous Symptom Vectors.

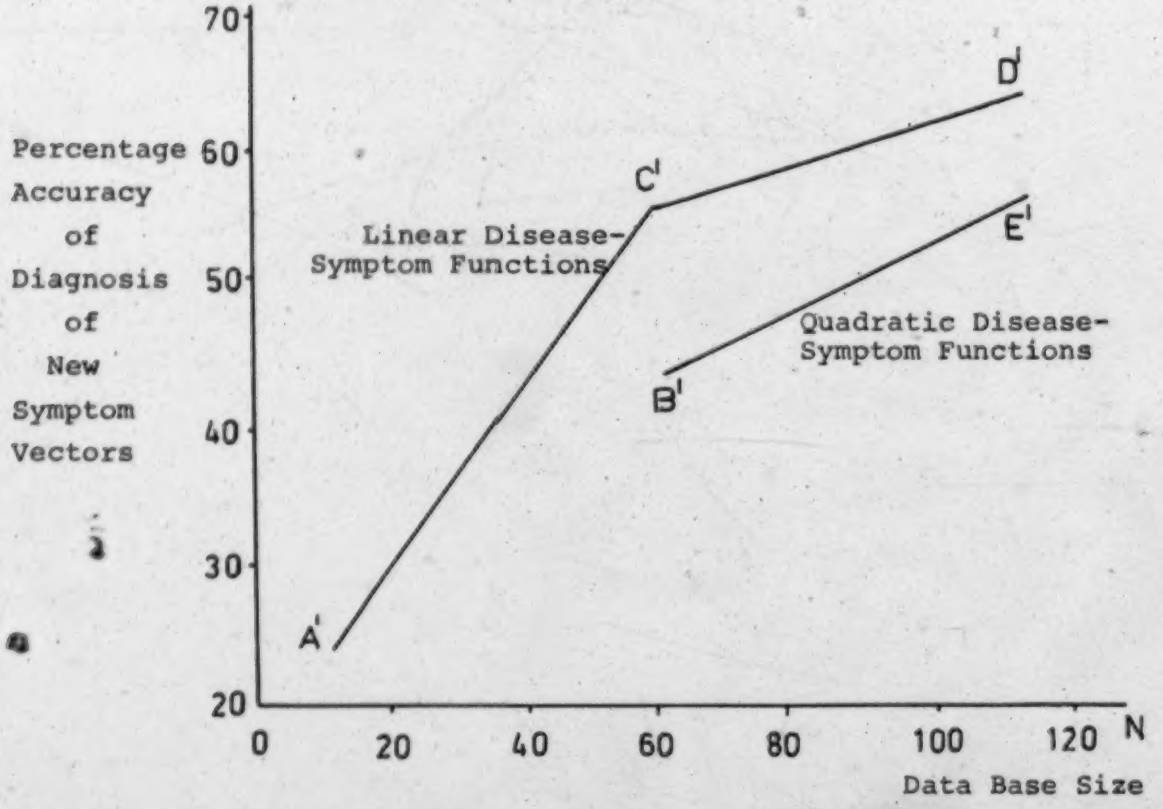

Figure 3.4 Percentage Accuracy of Diagnosis of New Symptom Vectors.

is increased. Figure 3.4 shows the accuracy of diagnosis of new symptom vectors using the corresponding linear (points A', C', D') and quadratic (points B', E') disease-symptom functions, as the data-base size N is increased. The graphs were obtained in the manner described in the following paragraphs.

From the reduced and modified data base 12 symptom vectors were selected such that the matrix $\bar{S}$ of (3.27) was non-singular. Accordingly the coefficients of the K linear disease-symptom functions were determined. The parameters β_k were set to zero and the α_k were chosen so that the $Z_k(n_{k1}) = 1$ and the $Z_k(n_{k2}) = 0$, giving point A in Figure 3.3.

These linear disease-symptom functions were then used to diagnose the 25 new symptom vectors. Only six such symptom vectors were correctly diagnosed, for an accuracy of 24% giving point A' in Figure 3.4.

The data base was then suitably increased so that the matrix $\bar{S}$ of (3.27) was non-singular when determining the coefficients of the K quadratic disease-symptom functions (3.32). When using the data supplied by Scheinok it is not possible to determine all 55 coefficients $C_{k\ell m}$, for the second symptom $S_2(n)$ is virtually redundant, occurring in 95% of all symptom vectors. Thus, in particular, nine of the symptom

pairs that involve $S_2(n)$ make the $\bar{S}$ matrix singular. These were removed, leaving a total of 58 coefficients to be determined.

The resulting quadratic disease-symptom functions then satisfy decision rule (3.2) for all 58 previous symptom vectors, shown by point B in Figure 3.3. The parameters β_k were again set to zero and the α_k chosen so that $Z_k(n_{k1}) = 1$ and $Z_k(n_{k2}) = 0$. The result of diagnosing the 25 new symptom vectors using these quadratic disease-symptom functions, was eleven correct classifications, shown by point B' in Figure 3.4.

The coefficients of the linear disease-symptom functions were then redetermined using the data base of size 58, for a correct diagnosis of 74% of the previous symptom vectors, giving point C in Figure 3.3. When these linear disease-symptom functions were applied to the new symptom vectors, the result was 16 correct diagnoses, giving point C' in Figure 3.4.

With the data base size increased up to its maximum size of 109 symptom vectors the coefficients of the linear and quadratic disease-symptom functions were redetermined. Diagnoses were made of both previous and new symptom vectors, giving points D, E and D', E' in Figures 3.3 and 3.4 respectively.

The graphs show that the accuracy of diagnosis of previous symptom vectors does not directly relate to

the accuracy of diagnosis of new symptom vectors. Further the transitions C to B and C' to B' show that the inclusion of the non-linear and interactive terms in (3.32) does not necessarily imply a higher accuracy of diagnosis of new symptom vectors.

The reason is that in diagnosing new symptom vectors, by reference to a data base of previous symptom vectors, it is necessary that the data base be representative of the new symptom vectors. As such the data bases of size 12, 58 and 109 were not representative. The increasing accuracy of diagnosis of new symptom vectors as the data base size increased indicates that the data base was becoming more representative. This suggests that, when using automatic methods of diagnosis, the data base should be as large as possible.

The lower accuracy obtained with quadratic disease-symptom functions when diagnosing new symptom vectors implies that the quadratic functions which satisfy the previous symptom vectors are not representative of those which satisfy the new symptom vectors. Herein lies the danger of using higher order non-linear and interactive terms in the disease-symptom functions. Points C' and D' show that a linear, rather than a quadratic function can be more representative of new symptom vectors.

3.8 Conclusion

It appears that the method developed in this chapter is suitable for automatic diagnosis. By appropriate choice of the parameters α_k and β_k, and inclusion of non-linear and interactive terms, the disease-symptom functions may be adjusted to suit any data base. The calculations are relatively simple, in that the major computation is the inversion of the matrix $(\bar{S} + \beta_k G_k)$. Using an IBM 360, Model 67, less than four minutes computation time were required to determine suitable α_k and β_k and to compute Table 3.1.

Once the coefficients of the disease-symptom functions have been determined, the diagnosis of any patient proceeds through K weighted-symptom summations. The decision rule (3.2) is then used to make the diagnosis. Such a procedure may be performed without difficulty even if a computing facility is not available.

CHAPTER 4

DISEASE PROBABILITIES

4.1 Introduction

The methods for automatic diagnosis reviewed in sections 2.5.2, 2.5.3 and 2.5.4, and as developed in Chapter 3 all use the weighted sum of the patient's symptoms for diagnosis. The decision rule, used with these methods, is such as to make the diagnosis definitive. Unfortunately this diagnosis is not always correct. Crooks diagnosed 15% of his patients incorrectly; Scheinok (1968) diagnosed 25% incorrectly. The alpha-beta method, using linear disease-symptom functions, diagnosed 19.7% of the previous patients incorrectly.

Since the resulting diagnosis is not always correct, it is appropriate to state the probability that the patient has the disease D_k. Then the statement, the patient p has the disease D_k, can be based upon the degree of certainty expressed by the probability. Bayesian methods do this, by using the symptoms to determine the probabilities

$$P(D_k | S_1(p) \ldots\ldots S_m(p) \ldots\ldots S_M(p)), \text{ all } k . \quad (4.1)$$

But, in doing so, the assumption is made that for each disease the symptoms are independent.

Disease-symptom functions make no assumptions as to symptom independence. Thus an alternative method of determining the probability that the patient p has the disease D_k is to use the values of the patient's disease-symptom functions. Further, all the advantages of disease-symptom functions, as outlined in section 3.1, are retained.

In this chapter, three different formulations of such disease probabilities are presented. Particularly the probabilities

$$P(D_k | z_1(p) \ldots . z_k(p) \ldots . z_K(p)), \text{ all } k, \qquad (4.2)$$

can be determined if it is assumed that, for each disease, the disease-symptom functions are independent.

Consideration is given to the problem of determining the coefficients of disease-symptom functions which are suitable for use with these disease probabilities. For two of the three formulations there is no known solution to this problem. In such instances the alpha-beta method can provide an approximate solution.

Such solutions have been found for the coefficients of the linear disease-symptom functions suitable for use in (4.2). Using data supplied by Scheinok (1972b) the resulting accuracy of diagnosis of

previous patients is found to be 82.5%. This compares with 74.9% when using (4.1) with the same data.

4.2 The Assumption of Normality

Several authors (Fisher, 1936; Sebestyen, 1962; Nilsson, 1965; Carl and Hall, 1972; Heaps, 1973; Cumberbatch, 1974) have formulated methods for determining the coefficients C_{km} of the linear disease-symptom function

$$Z_k(n) = \sum_m C_{km} S_m(n) \quad . \tag{4.3}$$

All have the common objective of separating the set

$$Z^{k1} = \{Z_k(n),\ n \in \Pi^{k1}\} \tag{4.4}$$

from the set

$$Z^{k2} = \{Z_k(n),\ n \in \Pi^{k2}\} \quad . \tag{4.5}$$

In (4.3) each $S_m(n)$ may be regarded as being a random variable sampled from a population. Hence each $Z_k(n)$ is a sum of random variables. Therefore the frequency distributions of Z^{k1} and of Z^{k2} are assumed to be normal.

If non-linear and/or interactive terms are introduced, new symptoms

$$S_{M+1} = S_1^2$$

$$S_{M+2} = S_1 . S_2 \quad , \quad \text{etc.}$$

may be defined as appropriate. The new symptoms S_{M+1}, S_{M+2}, etc. are also regarded as random variables, and the frequency distributions of Z^{k1} and of Z^{k2} are again assumed to be normal.

Accordingly the distribution of

$$\frac{Z_k(n) - E(Z_k(n))}{\sqrt{Var(Z_k(n))}} \qquad (4.6)$$

for $n \in \Pi^{k1}$ and for $n \in \Pi^{k2}$ will each be approximated by the normal distribution. All formulations of disease probabilities, as developed in this chapter, are based upon this assumption.

4.3 Disease Probabilities

In this section the probability $P(Z_k(p)|D_k)$ is determined by considering the value of $Z_k(p)$ in relation to the frequency distributions formed by Z^{k1} and Z^{k2}. Bayes' theorem is then used to determine the required probability $P(D_k|Z_k(p))$; i.e. the probability that the patient p has the disease D_k, given that the value of the k'th disease-symptom function, as determined from the patient's symptoms, is $Z_k(p)$.

This probability is considered to be limited in the sense that the only information used is that of the k'th disease-symptom function. In section 4.4 the probability is extended to use additional information.

Previous formulations for determining the coefficients C_k of the k'th disease-symptom function suitable for use with this probability are discussed. It is shown that the coefficients, as determined with the alpha-beta method, are suitable for use with this probability.

4.3.1 Limited Disease Probabilities

Let all previous patients not having the disease D_k be said to have the disease $D_{\bar{k}}$. Then since the diseases D_k and $D_{\bar{k}}$ are mutually exclusive

$$P(D_k|Z_k(n)) + P(D_{\bar{k}}|Z_k(n)) = 1.0 \ . \qquad (4.7)$$

Bayes' theorem provides the additional relation

$$\frac{P(D_k|Z_k(n))}{P(D_{\bar{k}}|Z_k(n))} = \frac{P(D_k)}{P(D_{\bar{k}})} \cdot \frac{P(Z_k(n)|D_k)}{P(Z_k(n)|D_{\bar{k}})} \ . \qquad (4.8)$$

The first term on the right of (4.8) is the ratio of the prior probabilities. The second quotient is called the "likelihood ratio". Let the probability density function of $Z_k(n)$, where $D(n) = D_k$, be denoted

by $f_k(Z_k(n))$, and let the probability density function of $Z_k(n)$, where $D(n) = D_{\bar{k}}$, be denoted by $f_{\bar{k}}(Z_k(n))$. Then since the $Z_k(n)$ are assumed to be continuous (by (4.6)) the likelihood ratio becomes

$$\frac{f_k(Z_k(n))}{f_{\bar{k}}(Z_k(n))} \tag{4.9}$$

(Van der Geer, 1971).

Accordingly it follows, from (4.7), (4.8) and (4.9), that the "limited disease probability" is given by

$$P(D_k | Z_k(n)) = \frac{f_k(Z_k(n)).P(D_k)}{f_k(Z_k(n)).P(D_k) + f_{\bar{k}}(Z_k(n)).P(D_{\bar{k}})} \tag{4.10}$$

and similarly

$$P(D_{\bar{k}} | Z_k(n)) = \frac{f_{\bar{k}}(Z_k(n)).P(D_{\bar{k}})}{f_k(Z_k(n)).P(D_k) + f_{\bar{k}}(Z_k(n)).P(D_{\bar{k}})} \tag{4.11}$$

The parameters needed for computation of (4.10) must all be estimated from the sets Z^{k1} and Z^{k2}. Let these sets have respective means $\bar{Z}_{k1}$ and $\bar{Z}_{k2}$, and standard deviations σ_{k1} and σ_{k2}. Then for the linear disease-symptom function (4.3)

$$\bar{Z}_{k1} = \bar{S}_k^T.C_k \tag{4.12}$$

$$\bar{Z}_{k2} = \bar{S}_{\bar{k}}^T.C_k \tag{4.13}$$

$$\sigma_{k1}^2 = \sum_{\ell} \sum_{m} g_{k\ell m} C_{k\ell} C_{km} = C_k^T G_k C_k \qquad (4.14)$$

and

$$\sigma_{k2}^2 = \sum_{\ell} \sum_{m} g_{\bar{k}\ell m} C_{k\ell} C_{km} = C_k^T G_{\bar{k}} C_k \qquad (4.15)$$

where

$$\bar{S}_{\bar{k}} = [\frac{1}{N_{k2}} \sum_{n_{k2}} S_m(n_{k2})] \qquad (4.16)$$

$M \times 1$

and

$$G_{\bar{k}} = [\frac{1}{N_{k2}} \sum_{n_{k2}} S_\ell(n_{k2}).S_m(n_{k2}) - \bar{S}_{k\ell}\bar{S}_{km}]. \qquad (4.17)$$

$M \times M$

Note that

$$\bar{S}_{\bar{k}} = \frac{1}{N_{k2}} \sum_{i \neq k}^{K} N_{i1}\bar{S}_i \qquad (4.18)$$

and

$$G_{\bar{k}} = \frac{1}{N_{k2}}(N\bar{S}-N_{k1}G_k-N_{k1}\bar{S}_k\bar{S}_k^T-N_{k2}\bar{S}_{\bar{k}}\bar{S}_{\bar{k}}^T) \qquad (4.19)$$

by (3.21), (3.23) and (4.17).

The prior probabilities may be estimated from the number of previous patients in each set Z^{k1} and Z^{k2} as N_{k1}/N and N_{k2}/N. Thus the required probability (4.10) is given by

$$P(D_k|Z_k(n)) = \frac{N_{k1}f_k(Z_k(n))}{N_{k1}f_k(Z_k(n)) + N_{k2}f_{\bar{k}}(Z_k(n))} \qquad (4.20)$$

where

$$f_k(Z_k(n)) = \frac{1}{\sqrt{2\pi\sigma_{k1}^2}} \exp -\left[\frac{(Z_k(n) - \bar{Z}_{k1})^2}{2\sigma_{k1}^2}\right] \qquad (4.21)$$

and

$$f_{\bar{k}}(Z_k(n)) = \frac{}{\sqrt{2\pi\sigma_{k2}^2}} \exp -\left[\frac{(Z_k(n) - \bar{Z}_{k2})^2}{2\sigma_{k2}^2}\right] . \qquad (4.22)$$

It is assumed that the new patients have symptoms $S_m(n^*)$ which follow the same particular distributions as those of all previous patients. Hence (4.20), (4.21) and (4.22) may be used to determine the disease probability of new patients.

The advantage of using disease probabilities lies in the decision criterion used to classify patients. For any level of confidence, viz. 95%, the patient p can only be said to have the disease D_k if,

$$P(D_k|Z_k(p)) \geq 0.95$$

and (4.23)

$$P(D_i|Z_i(p)) \leq 0.05 , \quad \text{all } i \neq k .$$

Further, the greater the value of $P(D_k|Z_k(p))$ and the smaller the value of the $P(D_i|Z_i(p))$, all $i \neq k$, the more likely is the diagnosis to be correct. This property is applied in the chapter on sequential diagnosis.

4.3.2 Formulations for Suitable Coefficients

Assume the validity of the assumption of normality. Then the limited disease probabilities $P(D_k|Z_k(p))$, all k, can always be evaluated, for any patient p, independently of the method used to determine the coefficients C_k of the functions $Z_k(p)$. However, the number of correct classifications that results from using the decision rule (4.23) is dependent upon the C_k.

It is possible to formulate the requirement of maximizing the number of correct classifications. But, with respect to determining the C_k, such formulations are invariably insoluble. However, by using some formulation which is representative of the decision rule used for classification, the coefficients C_k can be determined.

With this objective, it is possible to determine the C_k so as to maximize some measure of separation of z^{k1} and z^{k2}. If this separation is maximized the area of overlap under the frequency distribution curves formed by z^{k1} and z^{k2} is minimized. Hence the probability of classifying any $Z_k(n)$ as coming from z^{k1} (i.e. D_k) when in fact it is from z^{k2} (i.e. $D_{\bar{k}}$), or classifying any $Z_k(n)$ as coming from z^{k2} (i.e. $D_{\bar{k}}$) when in fact it is from z^{k1} (i.e. D_k) is minimized.

Greenhouse (1954) used this solution to the problem. He showed that the coefficients C_k, which

maximize Jeffrey's (1948) measure of separation, are of the form

$$C_k = a_1(b_1 G_k + G_{\bar{k}})^{-1}(\bar{S}_k - \bar{S}_{\bar{k}}) \qquad (4.24)$$

where a_1 and b_1 are scalars.

Greenhouse also considered Savage's (1954) measure of separation. The coefficients were again found to be of the form

$$C_k = a_2(b_2 G_k + G_{\bar{k}})^{-1}(\bar{S}_k - S_{\bar{k}}) \qquad (4.25)$$

where a_2 and b_2 are scalars.

In the instance that $G_k = G_{\bar{k}} = G$ both (4.24) and (4.25) are of the same form as that obtained by Fisher (1936), and by Anderson (1958), i.e.

$$C_k = a_3 G^{-1}(\bar{S}_k - \bar{S}_{\bar{k}}) \qquad (4.26)$$

where a_3 is scalar.

It is assumed that the probability density functions $f_k(Z_k(n))$ and $f_{\bar{k}}(Z_k(n))$ are normal, and that the $Z_k(n)$ are linear in the coefficients C_k. Hence the limited disease probability (4.20) is independent of any linear transformation of the coefficients C_k. Therefore all solutions (4.24), (4.25) and (4.26) can be written in the general form

$$C_k = (b\, G_k + G_{\bar{k}})^{-1}(S_k - \bar{S}_{\bar{k}}) \quad . \qquad (4.27)$$

The value of b is dependent upon the particular formulation used to determine the C_k.

4.3.3 Relation to the Alpha-Beta Method

The coefficients of the k'th linear disease-symptom function which are suitable for use with the limited disease probability have been shown to be of the form

$$C_k = (b\ G_k + G_{\bar{k}})^{-1}(\bar{S}_k - \bar{S}_{\bar{k}}) \qquad (4.28)$$

where b is a scalar. It is now shown that, under the assumption that the probability density functions $f_k(Z_k(n))$ and $f_{\bar{k}}(Z_k(n))$ are normal, the coefficients

$$C_k = \alpha_k(\bar{S} + \beta_k G_k)^{-1}\ \bar{S}_k\ , \qquad (4.29)$$

as obtained with the alpha-beta method, are of the same form as (4.28).

First note that the elements of the matrix $\bar{S}$ (in (4.29)) and given by

$$\bar{S}_{\ell m} = \frac{1}{N}\sum_n S_\ell(n).S_m(n)$$

$$= \frac{1}{N}[\sum_{n_{k1}} S_\ell(n_{k1})S_m(n_{k1}) + \sum_{n_{k2}} S_\ell(n_{k2})S_m(n_{k2})] \qquad (4.30)$$

so that, using the definitions of G_k (3.23) and $G_{\bar{k}}$ (4.17), it follows that

$$\bar{S} = \frac{1}{N}[N_{k1}G_k + N_{k2}G_{\bar{k}} + N_{k1}\bar{S}_k\bar{S}_k^T + N_{k2}\bar{S}_{\bar{k}}\bar{S}_{\bar{k}}^T] \quad . \tag{4.31}$$

For simplicity of notation let the suffices k_1 and k be denoted by 1, and the suffices k_2 and $\bar{k}$ be denoted by 2. Then

$$\begin{aligned}\bar{S} &= \frac{1}{N}[N_1G_1 + N_2G_2] + \frac{1}{N}[N_1\bar{S}_1\bar{S}_1^T + N_2\bar{S}_2\bar{S}_2^T] \\ &= \frac{1}{N}\sum_{i=1}^{2} N_iG_i + \frac{1}{N}\sum_{i=1}^{2} N_i\bar{S}_i\bar{S}_i^T \\ &= G + \frac{1}{N}\sum_{i}^{2} N_i\bar{S}_i\bar{S}_i^T \end{aligned} \tag{4.32}$$

where G is suitably defined.

For the particular instance that the redundant symptom $S_i(n) = 1$ is included, all n, the coefficients, as determined by the alpha-beta method, are given by the matrix equation

$$\underset{(M+1)\times(M+1)}{\begin{pmatrix} 1 & \left(\sum_{i}^{2}\frac{N_i}{N}\bar{S}_i\right)^T \\ \left(\sum_{i}^{2}\frac{N_i}{N}\bar{S}_i\right) & G+\beta_1G_1+\sum_{i}^{2}\frac{N_i}{N}\bar{S}_i\bar{S}_i^T \end{pmatrix}} \cdot \underset{(M+1)\times 1}{\begin{pmatrix} C_{10} \\ C_1 \end{pmatrix}} = \alpha_1 \underset{(M+1)\times 1}{\begin{pmatrix} 1 \\ \bar{S}_1 \end{pmatrix}} \tag{4.33}$$

This equation may be solved for C_{10} and C_1. First

$$C_{10} = \alpha_1 - \underset{1\times M}{\left(\sum_{i}^{2}\frac{N_i}{N}\bar{S}_i\right)^T}\ \underset{M\times 1}{C_1} \quad , \tag{4.34}$$

so that

$$\alpha_1\Big(\sum_i^2 \frac{N_i}{N}\bar{S}_i\Big) - \Big(\sum_i^2 \frac{N_i}{N}\bar{S}_i\Big)\cdot\Big(\sum_i^2 \frac{N_i}{N}\bar{S}_i\Big)^T C_1$$

$M\times 1 \qquad M\times 1 \qquad 1\times M \qquad M\times 1$

$$+ (G + \beta_1 G_1)\, C_1 + \Big(\sum_i^2 \frac{N_i}{N}\bar{S}_i\bar{S}_i^T\Big) C_1 = \alpha_1 \bar{S}_1 \qquad (4.35)$$

$M\times M \quad M\times 1 \qquad M\times M \quad M\times 1 \qquad M\times 1$

which becomes,

$$[(N-N_1)\,\bar{S}_1 - N_2\bar{S}_2]\Big[\frac{N_1}{N}\bar{S}_1 - \frac{N-N_2}{N}\bar{S}_2\Big]^T C_1$$

$M\times 1 \qquad\qquad 1\times M \qquad\qquad M\times 1$

$$+ N(G + \beta_1 G_1)\, C_1 = \alpha_1[(N-N_1)\,\bar{S}_1 - N_2\bar{S}_2] \quad . \qquad (4.36)$$

$M\times M \quad M\times 1 \qquad\qquad M\times 1$

Since $\big[\frac{N_1}{N}\bar{S}_1 - \frac{N-N_2}{N}\bar{S}_2\big]C_1$ is a scalar there exists some scalar X for which

$$[(N-N_1)\,\bar{S}_1 - N_2\bar{S}_2]\Big[\frac{N_1}{N}\bar{S}_1 - \frac{N-N_2}{N}\bar{S}_2\Big]^T C_1$$

$$= \alpha_1(1-X)[(N-N_1)\,\bar{S}_1 - N_2\bar{S}_2] \quad . \qquad (4.37)$$

Hence, substituting (4.37) into (4.36),

$$N(G + \beta_1 G_1)C_1 = \alpha_1 X[(N - N_1)\bar{S}_1 - N_2\bar{S}_2]$$

so that since $N - N_1 = N_2$

$$(G + \beta_1 G_1)C_1 = \frac{\alpha_1 X N_2}{N}\,(\bar{S}_1 - \bar{S}_2) \quad . \qquad (4.38)$$

Returning to the original suffices of k_1, k, k_2 and $\bar{k}$, and recalling that

$$G = \frac{1}{N} (N_{k1}G_k + N_{k2}G_{\bar{k}})$$

this gives

$$C_k = a(b\ G_k + G_{\bar{k}})^{-1}(\bar{S}_k - \bar{S}_{\bar{k}}) \qquad (4.39)$$

where

$$a = \alpha_k X\ , \qquad (4.40)$$

and

$$b = \frac{N_{k1} + \beta_k N}{N_{k2}}\ . \qquad (4.41)$$

Since the probability density functions $f_k(Z_k(n))$ and $f_{\bar{k}}(Z_k(n))$ are assumed to be normal, they are independent of the scalar a, and of the constant C_{ko} (4.34). Hence the coefficients, as determined by the alpha-beta method, are suitable for use with the limited disease probability. The value of β_k is dependent upon the particular formulation used to determine the C_k.

4.4 Extensions to Disease Probabilities

The preceding formulation of disease probability is based upon a suggestion by Duda and Hart (1973). In this section two further disease probabilities are formulated. The first of these extends Duda and Hart's suggestion by considering the frequency distribution of z^{k2} in the instance that $K > 2$. The second further extends the suggestion to use the information available

in the functions $Z_i(n)$, all $i \neq k$. The advantages of these extensions are discussed and it is argued that they result in a more accurate determination of disease probability.

The problem of determining the coefficients C_k which are suitable for use with these probabilities is too complex for conventional methods to be used. A constrained solution to this problem is found by using the coefficients obtained from the alpha-beta method. This solution uses the value of the parameter β_k which maximizes a measure of the separation of z^{k1} and z^{k2}. This measure is defined in terms of extensions to disease probabilities.

4.4.1 Extended Disease Probabilities

The limited disease probability (4.20) is determined by assuming that the frequency distribution of z^{k1} and of z^{k2} is normal. Actually the set z^{k2} is formed from the values of the $Z_k(n)$ of all previous patients having the K-1 diseases D_i, $i \neq k$. Thus, if $K > 2$, z^{k2} is a union of subsets;

$$z^{k2i} = \{Z_k(n),\ D(n) = D_i\}\ ,\quad i \neq k\ . \qquad (4.42)$$

Each $Z_k(n)$ in (4.42) is a sum of random variables, and it can therefore be assumed that the frequency distribution of each z^{k2i} is approximately normal. Thus

the limited disease probability can be extended to recognize that Z^{k2} is a union of subsets Z^{k2i}, $i \neq k$.

Since the diseases D_i, all $i \neq k$, are mutually exclusive (by (2.4)),

$$P(Z_k(n)|D_{\bar{k}}) = \sum_{i \neq k}^{K} \frac{P(Z_k(n).D_i)}{P(D_{\bar{k}})}$$

$$= \sum_{i \neq k}^{K} \frac{P(D_i).P(Z_k(n)|D_i)}{P(D_{\bar{k}})} \qquad (4.43)$$

(by Bayes' theorem).

Thus by substituting (4.43) into the denominator of (4.10) it can be shown that the "extended disease probability" is given by

$$P(D_k|Z_k(n)) = \frac{N_{k1} f_{kk}(Z_k(n))}{\sum_{i \leq K} N_{i1} f_{ik}(Z_k(n))} \qquad (4.44)$$

where $f_{ik}(Z_k(n))$ is the probability density function[1] of $Z_k(n)$, for $D(n) = D_i$. Thus

$$f_{ik}(Z_k(n)) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp -\left[\frac{(Z_k(n) - \bar{Z}_{ik})^2}{2\sigma_{ik}^2}\right] \qquad (4.45)$$

---

[1] The second subscript (k) of f_{ik} is introduced to make the notation consistent with that used to formulate the further extended disease probability (4.51).

where

$$\sigma_{ik}^2 = C_k^T G_i C_k \qquad (4.46)$$

and

$$\bar{z}_{ik} = \bar{s}_i^T C_k \quad . \qquad (4.47)$$

The extended disease probability (4.44) is thus inversely proportional to a weighted sum of the exponential functions (4.45). Hence the problem of determining the coefficients C_k which are suitable for use with this probability is mathematically complex. No solution is known.

4.4.2 Further Extended Disease Probabilities

In the instance that $K > 2$, additional information is available in the form of the other functions $Z_i(n)$, $i \neq k$. Thus the probability (4.43) can be further extended to the form

$$P(Z_1(n)....Z_k(n)....Z_K(n)|D_{\bar{k}})$$

$$= \sum_{i \neq k}^{K} \frac{P(D_i).P(Z_1(n)....Z_k(n)....Z_K(n)|D_i)}{P(D_{\bar{k}})} . \qquad (4.48)$$

By substituting (4.48) into a revised expression for (4.8) and using the relation

$$\sum_{i=1}^{K} P(D_i|Z_1(n)....Z_k(n)....Z_K(n)) = 1.0 \ , \qquad (4.49)$$

the "further extended disease probability"

$$P(D_k | Z_1(n) \ldots . Z_k(n) \ldots . Z_K(n))$$

can be determined. This is analogous to the conventional determination of disease probability, but using the functions $Z_k(n)$, for all k, rather than the symptoms $S_m(n)$, all m.

To evaluate (4.48) one of two assumptions must be made:- either that

$$P(Z_1(n) \ldots . Z_i(n) \ldots . Z_k(n) \ldots . Z_K(n) | D_i) \qquad (4.50)$$

can be determined from a multivariate normal density, or that for each disease D_i the $Z_k(n)$, all k, are independent.

Feller (1966) warns against the former assumption. He notes that the joint probability density of several variables is not necessarily normal even though the probability densities of each of these variables is normal.

The latter assumption can, however, be justified intuitively. Observe that it is the condition D_i, on the right of (4.50) which implies that $Z_i(n)$ is expected to fall within the range of all previous $Z_i(n)$ for which $D(n) = D_i$, and that the $Z_k(n)$, $k \neq i$, are expected to fall within the range of all previous $Z_k(n)$ for which $D(n) = D_i$, $i \neq k$. But, within these ranges of values, nothing can be said as to what values the $Z_k(n)$, all k, can be expected to have. This suggests that, for each

disease D_i, the $Z_k(n)$ may be assumed to be independent. Therefore

$$P(Z_1(n)....Z_k(n)....Z_K(n)|D_i)$$
$$= P(Z_1(n)|D_i)....P(Z_k(n)|D_i)....P(Z_K(n)|D_i),$$

and the further extended disease probability becomes

$$P(D_k|Z_1(n)....Z_k(n)....Z_K(n))$$
$$= \frac{N_{k1}f_{k1}(Z_1(n))...f_{kk}(Z_k(n))...f_{kK}(Z_K(n))}{\sum_{i\leq K} N_{i1}f_{i1}(Z_1(n))...f_{ik}(Z_k(n))...f_{iK}(Z_K(n))} \qquad (4.51)$$

where $f_{ik}(Z_k(n))$ for all i,k is given by (4.45).

The further extended disease probability (4.51) is dependent upon the coefficients C_k, all k. Hence the problem of determining coefficients which are suitable for use with this probability is extremely complex. No solution is known.

4.4.3 Advantages

There is no known solution to the problem of determining the coefficients C_k, of the k'th disease-symptom function, which are suitable for use with these extensions to disease probability. However, suppose that, for all k, the coefficients C_k are chosen so as to maximize some measure of separation of z^{k1} and z^{k2}.

Then, for the patient n, the values of the functions $Z_k(n)$, all k, are known. Thus, assuming the validity of the assumptions, the probabilities $P(D_k|Z_k(n))$, (4.44), and $P(D_k|Z_1(n)....Z_k(n)....Z_K(n))$, (4.51), can be determined.

But (4.44) and (4.51) recognize that Z^{k2} is a union of subsets. Thus (4.44) will determine the probability $P(D_k|Z_k(n))$ more accurately than will the limited disease probability (4.20).

In (4.51) the additional functions $Z_i(n)$, all $i \neq k$, are used. The coefficients C_i are chosen so that the resulting $Z_i(n)$, all n, maximize the separation of Z^{i1} and Z^{i2}. Since Z^{i1} is formed from the set $Z_i(n)$, where $n \in \Pi^{i1}$, for $i \neq k$, and Π^{i1} is a subset of Π^{k2} each $Z_i(n)$ provides some information as to the patient n having the disease $D_{\bar{k}}$. This is non-redundant information. Therefore, with respect to determining the probability that the patient n has the disease D_k, (using some function of the patient's symptoms) the probability (4.51) will be more accurate than will (4.44) or (4.20).

It is assumed that the new patients have symptoms $S_m(n^*)$ which follow the same particular distributions as those of all previous patients. Hence (4.44) and (4.51) may be used to determine the disease probabilities of new patients.

Note that the decision rule used to classify patients is simplified when using further extended disease probabilities (4.51). For if

$$P(D_k|Z_1(n)....Z_k(n)....Z_K(n)) \geq 0.95$$

then (4.49) ensures that

$$P(D_i|Z_1(n)....Z_k(n)....Z_K(n)) \leq 0.05, \text{ all } i \neq k .$$

Thus for any level of confidence, viz. 95%, any patient, p, can be said to have the disease D_k, if

$$P(D_k|Z_1(p)....Z_k(p)....Z_K(p)) \geq 0.95 . \qquad (4.52)$$

4.4.4 Determining Suitable Coefficients

It has been shown (section 4.3.3) that the coefficients, as obtained from the alpha-beta method, are suitable for use with limited disease probabilities (4.20). In the absence of any known solution to the problem of determining coefficients suitable for use with the disease probability extensions (4.44) and (4.51), and given that these probabilities can be determined using any coefficients, it is proposed that the coefficients used always be those obtained from the alpha-beta method;

$$C_k = (\bar{S} + \beta_k G_k)^{-1} \bar{S}_k . \qquad (4.53)$$

If $G_k = G_{\bar{k}}$, the expression (4.39) shows that the disease probabilities (4.44) and (4.51) are independent of β_k. However, if $G_k \neq G_{\bar{k}}$ the coefficients C_k are non-linearly dependent upon β_k. Suppose then that $\beta_k G_k$ is small in comparison to $\bar{S}$ (see (4.31)). Then a power series expansion of (4.53) may be truncated to give the linear approximation[2]

$$C_k = (I - \beta_k \bar{S}^{-1} G_k) \bar{S}^{-1} \bar{S}_k \quad . \tag{4.54}$$

Let the changes produced in $Z_k(n)$, $\bar{Z}_{ik}$ and σ^2_{ik}, by varying β_k from zero to a small non-zero value, be denoted by $\Delta Z_k(n)$, $\Delta \bar{Z}_{ik}$ and $\Delta \sigma^2_{ik}$. Substituting (4.54) into (4.3), (4.46) and (4.47) shows that

$$\Delta Z_k(n) = -\beta_k S(n) \bar{S}^{-1} G_k C_k \quad , \tag{4.55}$$

$$\Delta \bar{Z}_{ik} = -\beta_k \bar{S}_i^T \bar{S}^{-1} G_k C_k \quad , \tag{4.56}$$

and

$$\Delta \sigma^2_{ik} = -2\beta_k C_k^T G_i \bar{S}^{-1} G_k C_k \tag{4.57}$$

where the coefficients C_k are determined with $\beta_k = 0$. Both $\bar{S}^{-1}$ and G_k contain both positive and negative elements, and each of (4.55), (4.56) and (4.57) is unique. Therefore every disease probability (4.44) and (4.51), for all different symptom vectors $S(n)$, $n \in \Pi$, will be changed differently as β_k is varied.

Let the notation $P(D_k|Z(n))$ denote either probability (4.44) or (4.51). Then it is appropriate to choose β_k so that, for all $n_{k1} \in \Pi^{k1}$, $P(D_k|Z(n_{k1}))$ is maximized and $P(D_{\bar{k}}|Z(n_{k1}))$ is minimized. Further, for all $n_{k2} \in \Pi^{k2}$, β_k should be chosen so that $P(D_{\bar{k}}|Z(n_{k2}))$ is maximized and $P(D_k|Z(n_{k2}))$ is minimized.

It is intuitive to expect that these requirements can be met by maximization of such an expression as

$$J_k = \log \frac{[\frac{1}{N_{k1}} \sum_{n_{k1}} P(D_k|Z(n_{k1}))]}{[\frac{1}{N_{k1}} \sum_{n_{k1}} P(D_{\bar{k}}|Z(n_{k1}))]} - \log \frac{[\frac{1}{N_{k2}} \sum_{n_{k2}} P(D_k|Z(n_{k2}))]}{[\frac{1}{N_{k2}} \sum_{n_{k2}} P(D_{\bar{k}}|Z(n_{k2}))]} . \quad (4.58)$$

If each probability in (4.58) is determined from a normal density, then the denominator will be non-zero, and a finite maximum value of J_k will always exist. It is therefore proposed to use the β_k that maximize J_k.(2)

The expression J_k is based upon Jeffreys' and upon Kullback's (1968) measure of separation, but with the difference that the logarithms and summations have

---

(2) This is but one of several methods which could be used to determine suitable β_k. The advantage of using J_k is that J_k' (4.61) can be used as an approximation to J_k, and that J_k' is related to R_k (see (6.1)). Alternative methods, such as minimizing entropy or maximizing the number of correct classifications when using decision rules, do not have such advantage.

been interchanged. This rearrangement prevents J_k from being unduly influenced by a few small probabilities in the denominator. Unless otherwise indicated natural logarithms will be used.

The further extended disease probability (4.51) is dependent upon all K parameters β_k. Thus, when using (4.51), it is infeasible to determine the β_k which maximize J_k. However, the extended disease probability (4.44) is dependent upon only one parameter, β_k. Therefore, it is proposed that the chosen β_k be those which maximize each J_k (all k) as determined using (4.44). By this means suitable coefficients C_k (all k) can be found.

The value of β_k which maximizes each J_k is most sensibly found by searching from one estimate of β_k in the direction of another. Further, the distance between these two estimates provides a measure of the scale of the values of β_k.

A suitable first estimate is $\beta_k = 0$. The resulting coefficients are then of the form

$$C_k = G^{-1}(\bar{S}_k - \bar{S}_{\bar{k}}) \qquad (4.59)$$

(see (4.38)) where

$$G = \frac{1}{N}\,(N_{k1}G_k + N_{k2}G_{\bar{k}}) \quad . \qquad (4.60)$$

This is equivalent to using the coefficients determined

by Fisher and by Anderson where it is assumed that $\sigma_{k1}^2 = \sigma_{k2}^2 = C_k^T G C_k$, and G is determined from the pooled estimate (4.60), as proposed by Marascuilo (1971).

A second estimate can be found by determining that value of β_k for which $P(D_k|E(Z(n_{k1})))$ and $P(D_{\bar{k}}|E(Z(n_{k2})))$ are maximized, and $P(D_{\bar{k}}|E(Z(n_{k1})))$ and $P(D_k|E(Z(n_{k2})))$ are minimized. Using the same intuitive reasoning used to define J_k, the second estimate of β_k is that which maximizes the expression

$$J_k' = \log\left[\frac{P(D_k|E(Z(n_{k1})))}{P(D_{\bar{k}}|E(Z(n_{k1})))}\right] - \log\left[\frac{P(D_k|E(Z(n_{k2})))}{P(D_{\bar{k}}|E(Z(n_{k2})))}\right]. \quad (4.61)$$

In the instance that each probability in J_k' is of the form (4.10),

$$J_k' = \log\left[\frac{N_{k1} f_k(\bar{Z}_{k1})}{N_{k2} f_{\bar{k}}(\bar{Z}_{k1})}\right] - \log\left[\frac{N_{k1} f_k(\bar{Z}_{k2})}{N_{k2} f_{\bar{k}}(\bar{Z}_{k2})}\right]$$

$$= \frac{(\bar{Z}_{k1} - \bar{Z}_{k2})^2}{2\sigma_{k1}^2} + \frac{(\bar{Z}_{k1} - \bar{Z}_{k2})^2}{2\sigma_{k2}^2} \quad . \quad (4.62)$$

The usual calculus procedures determine the coefficients C_k of the k'th linear disease-symptom function which maximize (4.62) (see Appendix 2) as

$$C_k = a_4(b_4 G_k + G_{\bar{k}})^{-1}(\bar{S}_k - \bar{S}_{\bar{k}}) \quad (4.63)$$

where a_4 is scalar, and

$$b_4 = \left(\frac{C_k^T \bar{G}_k C_k}{C_k^T G_k C_k} \right)^2 \quad . \tag{4.64}$$

Note that this is consistent with the formulations for coefficients (4.24), (4.25) and (4.26).

If $\beta_k = 0$ is used as a first estimate of β_k, the coefficients C_k ($\beta_k = 0$) are known. Thus a second estimate of β_k (by (4.41)) is

$$\beta_k = \frac{N_{k2}}{N} \left(\frac{C_k^T(\beta_k=0) \bar{G}_k C_k(\beta_k=0)}{C_k^T(\beta_k=0) G_k C_k(\beta_k=0)} \right)^2 - \frac{N_{k1}}{N} \quad . \tag{4.65}$$

It is therefore proposed that the search for the β_k which maximize the J_k be initiated with $\beta_k = 0$, and proceed in the direction of β_k given by (4.65). If J_k increases, that direction is retained; if J_k decreases, the direction is reversed.

When the β_k which maximize the individual J_k (all k) have been found, so have the coefficients C_k, all k. All diagnoses should then be made using the further extended disease probabilities (4.51).

4.5 Results Using Disease Probabilities

The extensions to disease probabilities, as developed in this chapter, have been applied to the data supplied by Scheinok (1972b). It may be recalled that of the 300 previous patients there are 178 who

exhibit symptom vectors which are duplicated in two or more diseases. Any probabilistic method of diagnosis can, at best, determine the probability that each previous patient's symptom vector is associated with each disease. Therefore, to simplify the method of determining the resulting accuracy of diagnosis, and for purpose of comparison (see Chapter 5) with the results obtained in Chapter 3, the data base was again reduced to 223.

For further purpose of comparison the 223 previous patients were diagnosed using Bayes' theorem, on the assumption of symptom independence. As suggested by Scheinok (1967) the probabilities $P(D_k|S_m(n))$, all m, were determined using Bailey's correction for small samples.

All results presented are for the diagnosis of previous patients. The diagnosis of new patients is discussed in Chapter 6.

4.5.1 Assuming Normality

For this data the matrix G_k is not equal to the matrix $G_{\bar{k}}$, all k. Hence the procedure proposed in section 4.4.4 was used to determine the β_k which maximized each J_k, all k.

The second estimates of β_k were determined using (4.65). The values of β_k so obtained, to the nearest

multiple of 0.2, are shown in Table 4.1.

β_1	β_2	β_3	β_4	β_5	β_6
-0.2	0.8	0.2	0.4	0.4	0.4

Table 4.1

Second Estimates of β_k, to the Nearest Multiple of 0.2

In order to maximize each J_k, β_k was changed from zero, in increments of 0.2. The direction of the change was initially towards the corresponding β_k given in Table 4.1. If J_k increased, that direction was retained. If J_k decreased, the direction was reversed. The values of β_k so found to maximize each J_k are shown in Table 4.2.

β_1	β_2	β_3	β_4	β_5	β_6
-0.2	0.4	-0.2	0.6	0.4	-0.2

Table 4.2

Values of β_k Found to Maximize J_k, to the Nearest Multiple of 0.2

No attempts were made to find more accurate values of β_k. The value of J_k (all k) for the different values of β_k used in the search are shown in Table 4.3.

β_k	-0.4	-0.2	0.0	+0.2	+0.4
Hiatal Hernia k = 1	2.187	2.442	2.371	.	.
Gastric Ulcer k = 3	0.888	0.916	0.854	0.824	.
Functional Disease k = 6	1.191	1.273	1.166	1.141	1.138

β_k	0.0	+0.2	+0.4	+0.6	+0.8
Duodenal Ulcer k = 2	4.302	4.357	4.370	4.359	4.334
Cancer k = 4	4.630	5.217	5.374	5.410	5.409
Gallstones k = 5	4.538	4.570	4.572	4.568	.

Table 4.3 The Value of J_k for Different Values of β_k.

Tables 4.1 and 4.2 show that the initial direction of the search was correct more often than incorrect. Further, the second estimates of β_k are seen to be good. However, more applications are needed before this procedure can be fully evaluated.

Using values of $\beta_k = 0$, and those shown in Table 4.2 the extended disease probabilities (4.44) and further extended disease probabilities (4.51) of all 223 previous patients were determined. The accuracy of diagnosis was then based upon the number of correct classifications obtained using the decision rule:

$$\text{if } P(D_k|Z(n)) > P(D_i|Z(n)) , \quad \text{all } i \neq k$$

$$\text{then } D(n) = D_k .$$

Results are shown in Table 4.4. These results compare

	$\beta_k = 0$	$\beta_k \neq 0$
Extended disease probability	76.2	76.6
Further extended disease probability	75.4	76.2

Table 4.4

Percentage Accuracy of Diagnosis of 223 Previous Patients Using Disease Probabilities Assuming Normality.

with an accuracy of 74.9% obtained with Bayes' theorem, assuming symptom independence.

On reflection it was felt that the results shown in Table 4.4 are inconclusive. It was not expected that similar accuracies of diagnosis would be obtained. Reasons for these inconclusive results and consequent corrections are discussed in the next section.

4.5.2 Non-Normality

The extensions to disease probability have been formulated using the assumption that the frequency distribution fermed by each set Z^{ki}, all i,k, is normal. To test the validity of this assumption, as it applies to the data used herein, a frequency distribution analysis was performed on the particular sets Z^{kl}, all k. The results of this analysis, for two diseases, gallstones (k = 5) and functional disease (k = 6), are shown in Table 4.5. These two diseases were chosen as being representative of the data. The values of $Z_k(n)$ used are the same as those of Table 3.1 in Chapter 3.

For a normal distribution,mean, mode and median are equal, skewness is zero and kurtosis is three. The analysis revealed that no set Z^{kl}, all k, formed a normal distribution by these criteria.

Statistics on z^{kl} (k = 5)					
Mean	0.682	STD error	0.035	Median	0.652
Mode	0.652	STD dev	0.212	Variance	0.045
Kurtosis	0.311	Skewness	-0.453	Range	0.759
Minimum	0.231	Maximum	0.990		
Size of population = 37					

Statistics on z^{kl} (k = 6)					
Mean	0.370	STD error	0.029	Median	0.299
Mode	0.378	STD dev	0.160	Variance	0.026
Kurtosis	3.008	Skewness	-1.570	Range	0.752
Minimum	-0.265	Maximum	0.487		
Size of population. = 31					

Table 4.5 Frequency Distribution Analysis of z^{kl} (k=5,6).

The result of non-normality is that the probability density functions, used to determine the disease probabilities, are not exponential functions. Consequently all disease probabilities so determined are erroneous.

Consider again the extended disease probability

$$P(D_k|Z_k(n)) = \frac{P(D_k).P(Z_k(n)|D_k)}{\sum_{i\leq K} P(D_i).P(Z_k(n)|D_i)} \qquad (4.66)$$

and the further extended disease probability

$$P(D_k|Z_1(n)....Z_k(n)....Z_K(n))$$

$$= \frac{P(D_k).P(Z_1(n)|D_k)...P(Z_k(n)|D_k)...P(Z_K(n)|D_k)}{\sum_{i\leq K} P(D_i).P(Z_1(n)|D_i)...P(Z_k(n)|D_i)...P(Z_K(n)|D_i)} . \qquad (4.67)$$

Rather than assume normality, the conditional probabilities on the right of (4.66) and (4.67) can be determined from histogram presentations of the frequency distributions formed by Z^{ki}, all i,k.

To derive these histograms involves dividing the disease-symptom function space $(Z_k(n))$ into a finite number of intervals. The problem is to choose the intervals so that the resulting conditional probabilities are accurately estimated.

Hughes (1968) has plotted graphs to show the relation between the number of intervals used, per population size, and the percentage error incurred in estimating the probabilities. The curves show that the number of intervals becomes less critical as the population size increases.

For the data used in this study the smallest population size is for cancer, with $N_{kl} = 15$. For such a population the number of intervals should be from 2 to 4. For duodenal ulcer, with $N_{kl} = 72$, the number of intervals should be from 5 to 11.

In an attempt to meet these requirements, the functions $Z_k(n)$, all n,k, as determined with $\beta_k = 0$, were multiplied by 10, and rounded to the nearest integer. A computer program was then used to plot the frequency distributions formed by these transformed functions $Z_k'(n)$ in each disease, set Π^i, all i. Two of these distributions are shown in Figures 4.1 and 4.2. Inspection revealed that the number of intervals used satisfied the requirements for obtaining accurate estimates of the probabilities. Each conditional probability $P(Z_k(n)|D_i)$ was therefore determined from the frequency of occurrence of $Z_k'(n)$ in each disease set Π^i. Additionally Bailey's correction for small samples was used.

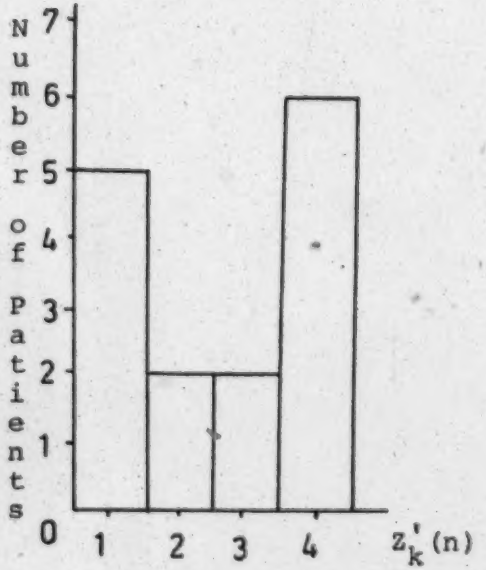

Figure 4.1 Histogram of Transformed $Z_k(n)$ for Patients Having Cancer.

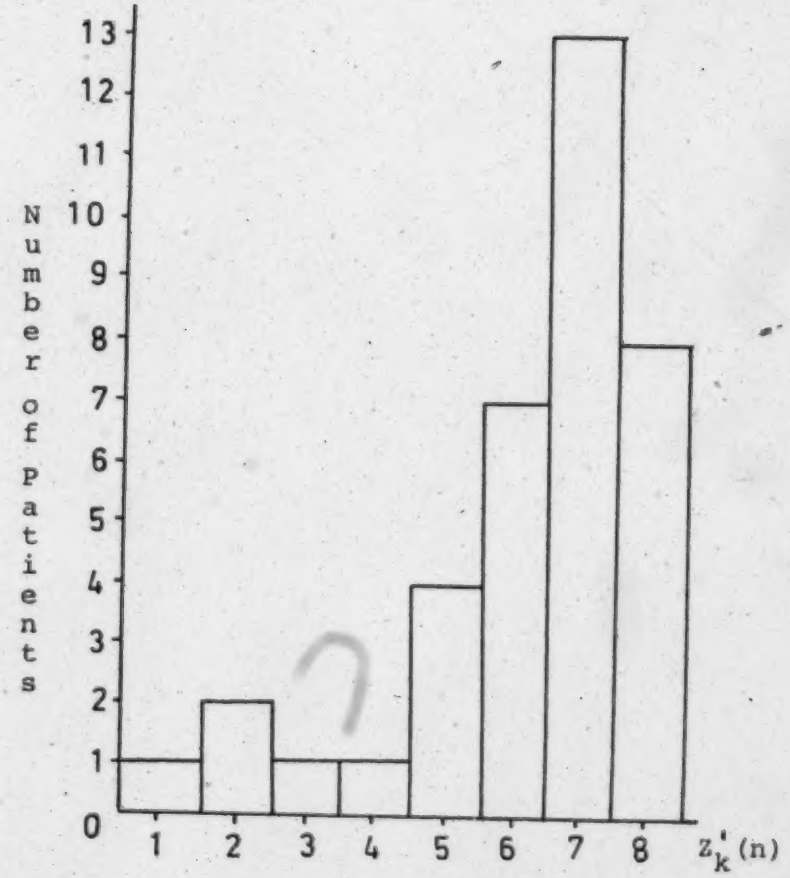

Figure 4.2 Histogram of Transformed $Z_k(n)$ for Patients Having Gallstones.

The values of β_k which maximized J_k under the assumption of normality were then used to determine the functions $Z_k(n)$, $(\beta_k \neq 0)$. Transformed functions $Z_k'(n)$ were again obtained and the resulting histograms plotted. The number of intervals used again satisfied the requirements for obtaining accurate estimates of the probabilities. Accordingly the conditional probabilities $P(Z_k(n)|D_k)$ were determined from the frequency of occurrence of $Z_k'(n)$ in each disease set Π^i. (Bailey's correction for small samples again being used).

The 223 previous patients were then diagnosed using extended disease probabilities (4.66) and further extended disease probabilities (4.67). Results are shown in Table 4.6. These results compare with an

	$\beta_k = 0$	$\beta_k \neq 0$
Extended disease probability	77.6	78.9
Further extended disease probability	79.4	82.5

Table 4.6

Percentage Accuracy of Diagnosis of 223 Previous Patients Using Disease Probabilities Not Assuming Normality.

accuracy of 74.9% obtained with Bayes' theorem, assuming symptom independence.

The results are consistent with expectation. If the additional information $Z_i(n)$, $i \neq k$, is used, then disease probability is more accurately determined.

Detailed results using further extended disease probabilities, and using Bayes' theorem, are presented in Tables 4.7 and 4.8. The "average information" shown in Table 4.7 was measured by evaluation of the expression $\sum_k \frac{-1}{N_{kl}} \sum_{n_{kl}} \log(P(D_k|Z(n_{kl})))$. This measure shows that the probability (4.67), in each of the K sets Π^{kl} was, on average, larger with $\beta_k \neq 0$ than with $\beta_k = 0$. It is therefore unlikely that the improved accuracy obtained with $\beta_k \neq 0$ was just chance.

4.5.3 Comment

The results presented in Table 4.7 would seem most attractive. An improvement of 7.6 in the percentage accuracy of diagnosis of previous patients is obtained compared with Bayes' theorem assuming symtom independence.

Unfortunately, all results overlook the errors incurred in estimating the required probabilities. Whether such errors could account for the whole improvement is not known. Rather it must be conceded that even if a similar improvement were to be obtained when diagnosing a random sample of 223 patients, the result would not be significant at the 95% confidence level (see section 3.7).

	Hiatal hernia	Duodenal ulcer	Gastric ulcer	Cancer	Gallstones	Functional disease	Totals	Percentage correct	Average information
Bayes	31	68	10	12	31	15	167	74.9	0.8833
Alpha-Beta									
$\alpha_k=1,\ \beta_k=0$	29	66	13	14	33	22	177	79.4	0.5577
$\alpha_k=1,\ \beta_k\neq 0$	35	63	15	15	32	24	184	82.5	0.5161
Number with disease	44	72	24	15	37	31	223		

Table 4.7 Number of Previous Patients Correctly Diagnosed When Using Further Extended Disease Probabilities Not Assuming Normality.

	Hiatal hernia	Duodenal ulcer	Gastric ulcer	Cancer	Gallstones	Functional disease	Totals	Percentage correct
Bayes	20	23	8	8	18	9	86	64.2
Alpha-Beta								
$\alpha_k=1,\ \beta_k=0$	18	22	10	10	20	15	95	70.9
$\alpha_k=1,\ \beta_k\neq 0$	22	22	11	11	19	15	100	74.6
Number with disease	31	27	20	11	23	22	134	

Table 4.8 Number of Symptom Vectors Correctly Associated with Their Disease When Using Further Extended Disease Probabilities Not Assuming Normality.

CHAPTER 5

COMPARISON OF RESULTS

5.1 Introduction

The accuracy of diagnosis of previous patients which resulted from use of the method developed in Chapter 3 was shown to be 80.3%. This compared with 70.8% using the least-squares-fit method proposed by Heaps. In Chapter 4 the further extended disease probabilities, as determined from histograms, resulted in an accuracy of diagnosis of previous patients of 82.5%. This compared with 74.9% using Bayes' theorem.

These accuracies of diagnosis were determined from the number of correct classifications. In this chapter consideration is given to the manner in which each method divides up the disease-symptom space into regions for classification. By this means explanation can be given as to why higher accuracies of diagnosis were obtained using the methods developed in this thesis.

5.2 Linear Separating Surfaces

In Chapter 3 the patient n was classified as having the disease D_k if $Z_k(n) > Z_i(n)$, all $i \neq k$. The region for the classification $D(n) = D_k$ was

therefore bounded by the separating surfaces given by

$$Z_k(n) - Z_i(n) = 0, \quad \text{all } i \neq k . \tag{5.1}$$

Consider then the separating surface

$$Z_k(n) - Z_i(n) = 0, \qquad i \neq k . \tag{5.2}$$

When using Heaps' method

$$Z_k(n) = \sum_{m=1}^{M} C_{km} S_m(n) \tag{5.3}$$

so that (5.2) is given by

$$\sum_{m=1}^{M} S_m(n)(C_{km} - C_{im}) = 0 . \tag{5.4}$$

When using the alpha-beta method

$$Z_k(n) = \alpha_k \left[\sum_{m=1}^{M} C_{km}(\beta_k) S_m(n) + C_{ko}(\beta_k) \right] \tag{5.5}$$

so that (5.2) is given by

$$\sum_{m=1}^{M} S_m(n)[\alpha_k C_{km}(\beta_k) - \alpha_i C_{im}(\beta_i)]$$

$$+ \alpha_k C_{ko}(\beta_k) - \alpha_i C_{io}(\beta_i) = 0 . \tag{5.6}$$

Thus (5.6) avoids the restriction that all separating surfaces pass through the origin. Further, (5.6) permits the parameters α_k, α_i, β_k, and β_i to be varied in an attempt to find that separating surface which results in the maximum number of correct classifications.

In (5.4) the separating surface is rigidly defined by the coefficients C_{km}, C_{im}.

In Chapter 4 the patients were classified according to the most probable disease. Hence, when using Bayes' theorem, as applied to the symptoms, (5.2) becomes

$$P(D_k|S_1(n).....S_M(n)) - P(D_i|S_1(n).....S_M(n)) = 0. \quad (5.7)$$

Assuming symptom independence, (5.7) is satisfied when

$$\frac{P(D_k).P(S_1(n)|D_k).....P(S_M(n)|D_k)}{P(D_i).P(S_1(n)|D_i).....P(S_M(n)|D_i)} = 1 \qquad (5.8)$$

or

$$\sum_{m=1}^{M} \log \frac{P(S_m(n)|D_k)}{P(S_m(n)|D_i)} + \log \frac{P(D_k)}{P(D_i)} = 0 . \qquad (5.9)$$

For the data used in this application, the symtoms were binary valued. Let b_{km} denote the probability that $S_m(n) = 1$ in the disease set Π^{kl}. Then

$$P(S_m(n)|D_k) = b_{km}^{S_m(n)} . (1 - b_{km})^{1-S_m(n)} . \qquad (5.10)$$

Substituting (5.10) into (5.9) gives

$$\sum_{m=1}^{M} S_m(n) \log\left(\frac{b_{km}(1 - b_{im})}{b_{im}(1 - b_{km})}\right) + P_o = 0 \qquad (5.11)$$

where P_o is a constant.

Inspection shows that (5.11) is linear in the $S_m(n)$ and is of the same form as (5.6). Therefore it is valid to compare the results of Chapter 3 (80.3% with the alpha-beta method) with those obtained in Chapter 4 using Bayes' theorem as applied to the symptoms (74.9%). The reason for the lower accuracy when using Bayes' theorem is the assumption of symptom independence which constrains the resulting separating surface to the form (5.11) defined by the b_{km}, b_{im} and P_o.

Disease sets can be linearly separated even when the symptoms are not independent. Presumably (5.11) would not result in such a separation, whereas (5.6) would.

For binary valued symptoms the symptom vectors S(n), all n, appear as the vertices of an M dimensional hypercube. The surfaces (5.4), (5.6) and (5.11) are hyperplanes which attempt to separate the $D(n) = D_k$ vertices from the $D(n) = D_i$ vertices. For this data the results showed that the best separating hyperplanes found were of form (5.6), which resulted in the misclassification of 19.7% of the vertices. However, unless the $D(n) = D_k$ vertices are linearly separable from the $D(n) = D_i$ vertices, all $i \neq k$, then no hyperplanes can be found which will result in the correct

classification of all the $D(n) = D_k$ vertices. Thus the upper bound on the accuracy of diagnosis when using (5.4), (5.6) and (5.11) is not necessarily 100%. Perhaps it may be argued that given the flexibility of varying α_k and β_k, all k, the 80.3% accuracy so obtained is close to the upper limit obtainable when using separating hyperplanes.

5.3 Non-Linear Separating Surfaces

Turning now to the further extended disease probabilities used in Chapter 4, the resulting separating surface is given by

$$\sum_{j=1}^{K} \log \frac{P(Z_j(n)|D_k)}{P(Z_j(n)|D_i)} + \log \frac{P(D_k)}{P(D_i)} = 0 . \qquad (5.12)$$

If normal probability densities are assumed (5.12) becomes

$$\sum_{j=1}^{K} \left(\frac{(Z_j(n)-\bar{Z}_{ij})^2}{2\sigma_{ij}^2} - \frac{(Z_j(n)-\bar{Z}_{kj})^2}{2\sigma_{kj}^2} + \log \frac{\sigma_{ij}^2}{\sigma_{kj}^2} \right) + \log \frac{P(D_k)}{P(D_i)} = 0 . \qquad (5.13)$$

When using linear disease-symptom functions each separating surface (5.13) is a hyperquadratic. Although a hyperquadratic is superior to a hyperplane when used as a separating surface, the assumptions upon which (5.13) is based were shown to be false. Hence the

method performed worse (76.2%) than when using the hyperplanes (5.6) (80.3%).

If the probabilities $P(Z_j(n)|D_k)$, all j,k, are determined from histograms then, for any given value of $Z_j(n)$, a number X_j can always be found such that

$$\frac{P(Z_j(n)|D_k)}{P(Z_j(n)|D_i)} = X_j^{Z_j(n)} \quad . \quad (5.14)$$

Hence (5.12) becomes

$$\sum_{j=1}^{K} Z_j(n) \log X_j + \log \frac{P(D_k)}{P(D_i)} = 0 \ , \quad (5.15)$$

which, when using linear disease-symptom functions, becomes

$$\sum_{m=0}^{M} S_m(n) [\sum_{j=1}^{K} C_{jm} \log X_j] + \log \frac{P(D_k)}{P(D_i)} = 0. \quad (5.16)$$

As the $S_m(n)$ are varied, to map the separating surface, each X_j will change whenever $Z_j(n)$ moves into a new interval along the histogram's Z_j axis. Thus a piecewise linear separating surface results.

If the probabilities in (5.14) are unbiased then (5.16) will not be over-determined. Then if the coefficients C_{jm}, all j, are suitably chosen the piecewise linear separating surfaces will likely result in more previous patients being correctly classified (82.5%) than will the hyperplane (5.6) (80.3%).

CHAPTER 6

SEQUENTIAL DIAGNOSIS

6.1 Introduction

In the preceding chapters the diagnosis of any patient p was made only when the entire set of symptoms S(p) were known. However, in many instances the determination of the value of the symptoms may involve not only considerable expense, but also discomfort and may be hazard to the patient. Hence, it may be preferable to attempt a diagnosis with a limited set of symptoms.

Indeed, recent studies would indicate that a limited set of symptoms will often suffice. Pipberger (1968) observed that at certain hospitals each patient suffering from chest pain was asked 429 questions, and subjected to 69 tests. Yet, after using discriminant function analysis on each symptom, Pipberger was able to show that with fewer than 10 of the symptoms[1] more than 95% of 1000 additional patients suffering from either coronary artery disease or pneumonia could be correctly diagnosed.

---

[1] Again "symptom" is used as a generalization to include signs and test results.

If a diagnosis made with a reduced set of symptoms is too indefinite then additional symptoms need to be incorporated. It is the object of sequential diagnosis to determine which symptom should be chosen next, according to prescribed criteria.

Gorry (1968) used a symptom-selection function which combined the cost of determining the value of each additional symptom with the resulting "cost" of misdiagnosis. The cost of determining each symptom was taken to be 1.0, and the cost of every possible misdiagnosis was taken to be 1000. The method was applied to the sequential diagnosis of patients suffering from 35 different congenital heart diseases. It was found that on average only 6.9 symptoms were needed to obtain results comparable with those of expert clinicians using 34 symptoms.

Taylor (1972) used entropy as the criterion for selecting additional symptoms. This measure makes it possible to determine which symptom can be expected to yield the most information at the current stage in the diagnosis. Taylor compared this method with another which took additional account of the financial costs incurred in determining the value of the additional symptoms. The cost-conscious method was found to be as accurate as the cost-free method and in 67 cases of

thyroid enlargement was 30% cheaper. Both methods used only one third of the full set of symptoms.

Both Gorry and Taylor used the current diagnosis to decide which symptom to choose next. Gleser (1972) has suggested that the alternative approach is to use the data base to select additional symptoms in a sequence which leads to the correct diagnosis of all previous patients with respect to their having, or not having, any specified disease. Such a sequence is particularly convenient for screening purposes, where it is required that a common sequence of tests be applied to all patients, in an attempt to determine whether or not they have a specified disease.

Gleser used average entropy to measure the uncertainty that the previous patients have, or have not, a specified disease. By this means each additional symptom can be assigned a number equal to the reduction in uncertainty which will be obtained by choosing that symptom. The properties of this measure permit the determination of the probability that such a decrease in uncertainty could have occurred by chance alone. The next symptom chosen in the sequence was that having the lowest such probability. The method was used to develop a sequence of symptoms which would lead to the

correct diagnosis of previous patients having and not having "unrecognized" diabetes mellitus, the objective being that any future patient diagnosed as having "unrecognized" diabetes mellitus would be given a glucose tolerance test prior to seeing the doctor.

In this chapter the "diagnostic value" of each symptom is the measure of the extent to which the accuracy of diagnosis of previous patients is changed by the addition of that symptom. The cost of determining the value of each symptom may be included by choosing, as the next symptom, that having the largest diagnostic value per unit cost.

It is shown that the diagnostic value, so defined, is disease conscious; that is, it is likely to be different with respect to different diseases. A method is proposed for determining the diagnostic value of any additional symptom with respect to several diseases. Hence the current most probable disease may be used in deciding which diseases should be considered when selecting the next symptom. Alternatively the method may be used to determine the sequence of symptoms which will lead to the diagnosis of any patient as having or not having any specified disease.

The method is applied to new patients in a presumed environment where the doctor required confirmation, by selection of additional symptoms, of the current

diagnosis. A comparison of a disease-conscious and a non-disease-conscious selection of additional symptoms shows that the former can confirm the diagnosis using fewer symptoms than the latter.

6.2 The Diagnostic Value of a Symptom

Suppose that the diagnosis of any patient p is made using a limited set of symptoms. Then the decision whether or not to accept the current diagnosis as definitive can be made from knowledge of the probability that the patient p has each disease D_k, all k.

If it is decided to select an additional symptom, criteria for symptom selection are needed. A suitable criterion is to choose that symptom of largest diagnostic value, as measured by the extent to which the disease probabilities of previous patients are improved by addition of that symptom. It is assumed that a corresponding improvement will be obtained in the disease probabilities of the patient p.

An improvement in the disease probabilities of the previous patients is meant to imply an increase in diagnostic accuracy. Hence such a diagnostic value is a measure of the extent to which the accuracy of diagnosis of previous patients is increased by the addition of that symptom.

6.2.1 With Respect to One Disease

The set Π, composed of all previous patients, can be partitioned into two sets Π^{k1} and Π^{k2}. Hence the diagnostic value of any symptom, with respect to the k'th disease, can be measured by the improvement in the disease probabilities of patients in the set Π^{k1} and Π^{k2} that results by addition of that symptom.

A suitable measure of the disease probabilities in the sets Π^{k1} and Π^{k2} is J_k, as defined in (4.58). However, in order to evaluate J_k, the probability that every previous patient has the disease D_k and also $D_{\bar{k}}$ must be determined. Thus in order to minimize the computation time it is necessary to make certain simplifying assumptions.

First consider the measure J_k' (4.61) as being an approximation to J_k. Then, if the limited disease probability (4.10) is used and the additional simplifying assumption is made that $\sigma_{k1}^2 = \sigma_{k2}^2 = C_k^T G C_k$ it can be shown (see Appendix 3) that R_k $(\beta_k = 0)$ is directly related to J_k' by the relation

$$R_k^2(\beta_k=0) = 1 + \frac{N_{k2}^2}{N_{k1}N_{k2} + \frac{N^2}{J_k'}} \quad . \tag{6.1}$$

Hence the diagnostic value of an additional symptom, with respect to the k'th disease, may be measured by

the extent to which the addition of that symptom increases the value of

$$R_k^2(\beta_k = 0) = \frac{\bar{Z}_{k1}}{\frac{1}{N}\sum_n Z_k^2(n)} = \frac{C_k^T \bar{S}_k . \bar{S}_k^T C_k}{C_k^T \bar{S} C_k}$$

$$= \frac{\bar{S}_k^T \bar{S}^{-1} \bar{S}_k . \bar{S}_k^T \bar{S}^{-1} \bar{S}_k}{\bar{S}_k^T \bar{S}^{-1} \bar{S} \bar{S}^{-1} \bar{S}_k} = \bar{S}_k^T \bar{S}^{-1} \bar{S}_k \ . \qquad (6.2)$$

Viewed in this manner the diagnostic value of a symptom depends on the set of previous symptoms as well as on the symptom itself.

The relation (6.1) reveals a consistency between disease-symptom functions, disease probabilities (assuming normality) and diagnostic value. Further by using $R_k^2(\beta_k = 0)$ as a measure of diagnostic value all the advantages of disease-symptom functions are retained.

Consider the particular instance in which linear disease-symptom functions are used. Then inclusion of an additional symptom, designated the q'th symptom, extends the $\bar{S}$ matrix of (6.2) by the addition of a row of elements $\bar{S}_{q1}, \bar{S}_{q2}, \ldots, \bar{S}_{qM}, \bar{S}_{qq}$ and a column of elements $\bar{S}_{1q}, \bar{S}_{2q}, \ldots, \bar{S}_{Mq}, \bar{S}_{qq}$. The extended $\bar{S}$ matrix can therefore be written in the form

$$\bar{S}^* = \begin{pmatrix} \bar{S} & R^T \\ R & \bar{S}_{qq} \end{pmatrix}$$

$$(M+1) \times (M+1)$$

where

$$R = (\bar{S}_{q1}, \bar{S}_{q2}, \ldots, \bar{S}_{qm}, \ldots \bar{S}_{qM}) \quad . \tag{6.3}$$

A bordering expansion of the $\bar{S}$ matrix shows that the inverse of $\bar{S}^*$ is given by

$$\bar{S}^{*-1} = \begin{pmatrix} V & U^T \\ U & 1/w_{qq} \end{pmatrix} \tag{6.4}$$

$(M+1)\times(M+1)$

where

$$\underset{1\times M}{U} = - \underset{1\times M}{(R} \underset{M\times M}{\bar{S}^{-1}})/w_{qq} \tag{6.5}$$

$$\underset{M\times M}{V} = \underset{M\times M}{\bar{S}^{-1}} + \underset{M\times 1}{(R \bar{S}^{-1})^T} \underset{1\times M}{(R \bar{S}^{-1})}/w_{qq} \tag{6.6}$$

and

$$w_{qq} = \bar{S}_{qq} - \underset{1\times M}{(R \bar{S}^{-1})} \underset{M\times 1}{R^T} \quad . \tag{6.7}$$

Hence, putting

$$\underset{1\times M}{t} = R \bar{S}^{-1} = [\sum_{i}^{M} \underset{1\times M}{\bar{S}_{qi}} d_{ij}] \ , \tag{6.8}$$

where d_{ij} are the elements of $\bar{S}^{-1}$, and denoting the resulting value of R_k by R_{kq}, substitution into (6.2) shows that

$$R^2_{kq} = \bar{S}_k \bar{S}^{-1} \bar{S}_k + \frac{\bar{S}^T_k t^T \bar{S}_k - \bar{S}^T_k t^T \bar{S}_{kq} - \bar{S}_{kq} t\, \bar{S}_k + \bar{S}^2_{kq}}{w_{qq}}$$

$$= R^2_k + \frac{(\bar{S}_{kq} - t\, \bar{S}_k)^2}{w_{qq}} \quad . \tag{6.9}$$

The expression

$$\Delta R^2_{kq} = \frac{(\bar{S}_{kq} - t\, \bar{S}_k)^2}{w_{qq}}$$

$$= \frac{(\bar{S}_{kq} - \sum_i \sum_j \bar{S}_{qi} d_{ij} \bar{S}_{kj})^2}{(\bar{S}_{qq} - \sum_i \sum_j \bar{S}_{qi} d_{ij} \bar{S}_{qj})} \tag{6.10}$$

is the diagnostic value of the q'th symptom with respect to the k'th disease.

The addition of the q'th symptom modifies the expression for R_k by the addition of a further coefficient C_{kq}. Since the value of R_k is independent of the order of the coefficients

$$R^2_{kqr} = R^2_k + \Delta R^2_{kq} + \Delta R^2_{kqr}$$

$$= R^2_k + \Delta R^2_{kr} + \Delta R^2_{krq} = R^2_{krq} \tag{6.11}$$

where the q'th and r'th symptoms are added in the order implied by their ordering as subscripts.

Let the cost of determining the q'th symptom be e_q. Then, by considering the ratio $\Delta R^2_{kq}/e_q$, the next symptom chosen may be that having the highest diagnostic value per unit cost.

It should be noted that (6.10) can be reduced to an indefinite. Suppose that the q'th symptom is totally redundant; then for some scalar a,

$$S_q(n) = a\, S_h(h) \quad , \quad \text{for all } n \in \Pi,$$

and for some $0 \leq h \leq M$. Then since

$$\sum_i \bar{S}_{hi} d_{ij} = \begin{cases} 1 , & h = j \\ 0 , & h \neq j \end{cases} \tag{6.12}$$

it follows that

$$\sum_i \sum_j \bar{S}_{qi} d_{ij} \bar{S}_{qj} = a^2 \sum_i \sum_j \bar{S}_{hi} d_{ij} \bar{S}_{hj}$$

$$= a^2 \bar{S}_{hh} = \bar{S}_{qq} \tag{6.13}$$

and similarly

$$\sum_i \sum_j \bar{S}_{qi} d_{ij} \bar{S}_{kj} = \bar{S}_{kq} \, . \tag{6.14}$$

Thus ΔR^2_{kq} is reduced to an indefinite, reflecting the occurrence of singularity in the $\bar{S}^*$ matrix. Such symptoms are of no value for diagnosis and may be disregarded.

Alternatively suppose that the q'th symptom is the first symptom chosen. Suppose further that this q'th symptom is a perfect discriminant for patients having the disease D_k. Then by inclusion of the symptom $S_o(n) = 1$, all $n \in \Pi$, the quantization of the q'th

symptom may be chosen so that

$$S_q(n) = \bar{S}_q \ , \quad \text{all } n \in \Pi^{k1}$$

and

$$S_q(n) = 0 \ , \quad \text{all } n \in \Pi^{k2} \ .$$

Then

$$\Delta R^2_{kq} = \frac{(\bar{S}_{kq} - \bar{S}_{qo}\bar{S}_{ko})^2}{\bar{S}_{qq} - \bar{S}_{qo}\bar{S}_{qo}} \tag{6.15}$$

where

$$\bar{S}_{kq} = \bar{S}_q \ , \qquad S_{qo} = N_{k1}\bar{S}_q/N \ ,$$

$$\bar{S}_{qq} = N_{k1}\bar{S}^2_q/N \ , \qquad \bar{S}_{ko} = 1 \ ,$$

so that

$$\Delta R^2_{kq} = \frac{N - N_{k1}}{N_{k1}} \ . \tag{6.16}$$

Hence

$$R^2_{kq} = R^2_k + \Delta R_{kq} = 1 + \frac{N - N_{k1}}{N_{k1}} = \frac{N}{N_{k1}} \tag{6.17}$$

which is the upper bound of R^2_k.

The additional symptom S_q extends the k'th linear disease-symptom function by the addition of a term $C_{kq}S_q(n)$. With $\beta_k = 0$ the coefficients C_q are chosen so as to maximize $Q_k = R_k$. Since the additional coefficient C_{kq} may be set to zero, it follows that any non-zero value of C_{kq} will necessarily correspond to a further maximization of R_k. Accordingly the diagnostic value of

any non-redundant symptom, S_q, is always in the range

$$0 < \Delta R_{kq} \leq \frac{N - N_{k1}}{N_{k1}} \quad . \qquad (6.18)$$

In the instance that quadratic, cubic, etc., disease-symptom functions are used, the matrices $\bar{S}$ and $\bar{S}_k$ of (6.2) can be expanded to include all appropriate elements. Hence the diagnostic value of the q'th symptom with respect to the k'th disease may still be measured by the extent to which

$$R_k^2 = \bar{S}_k^T . \bar{S}^{-1} . \bar{S}_k$$

is increased by the addition of the q'th symptom.

6.2.2 With Respect to Several Diseases

The diagnostic value (6.10) of the q'th symptom with respect to the k'th disease is dependent upon the terms $\bar{S}_{kj}$, all $j \leq M$, and $\bar{S}_{kq}$ and it has an upper bound (6.18) dependent upon N_{k1}. Since $\bar{S}_{kj}$, $j \leq M$, $\bar{S}_{kq}$ and N_{k1} are likely to be different for each $k \leq K$, it follows that (6.10) will also likely be different for each $k \leq K$.

If each ΔR_{kq}^2 of form (6.10) is normalized, the diagnostic value of the q'th symptom with respect to L different diseases may be obtained by a summation of all L such terms. Therefore the expression

$$\Delta R^2_{Lq} = \sum_{k}^{L} \frac{N_{k1} \cdot \Delta R^2_{kq}}{(N - N_{k1})} \qquad (6.19)$$

is the diagnostic value of the q'th symptom with respect to L different diseases.

Each ΔR^2_{kq} in (6.19) is a measure of the extent to which the accuracy of diagnosis of previous patients, with respect to their having or not having the disease D_k, is increased by the addition of the q'th symptom. Hence in the particular instance that L = K, the measure ΔR^2_{Kq} may be used to determine the symptom which gives the greatest improvement in the accuracy of diagnosis of previous patients having the disease D_k, all $k \leq K$.

6.3 The Diagnostic Value of Several Symptoms

The sequential inclusion of additional symptoms increases the accuracy of diagnosis. However, if the symptoms are included one at a time, the number of symptoms used to reach any specific level of accuracy is not necessarily minimal.

As an illustrative example consider the following case involving six patients, three diseases (D_1, D_2 and D_3) and three symptoms (S_1, S_2 and S_3). The patient records (2.5) are

n ;	$S_1(n)$	$S_2(n)$	$S_3(n)$;	D(n)
1 ;	1	1	1 ;	1
2 ;	1	1	1 ;	1
3 ;	0	1	0 ;	2
4 ;	0	1	0 ;	2

n ;	$S_1(n)$	$S_2(n)$	$S_3(n)$;	D(n)
5 ;	1	0	1 ;	2
6 ;	1	1	0 ;	3
7 ;	0	0	1 ;	3
8 ;	0	0	1 ;	3

The diagnostic value of S_1 with respect to the first disease, D_1, is greater than that of S_2 or S_3 (by 6.15). Thus if the symptoms are chosen one at a time S_1 is chosen first. But if S_1 is chosen first the correct diagnosis of all the previous patients, with respect to their having or not having D_1, can only be obtained when S_1, S_2 and S_3 are used. However if the symptoms are chosen two at a time then the required correct diagnosis (D_1 versus $D_{\bar{1}}$) is obtained using S_2 and S_3.

The expression

$$\Delta R^2_{kq} = \frac{(\bar{S}_{kq} - t.\bar{S}_k)^2}{w_{qq}}$$

is restricted to determining the diagnostic value of symptoms chosen one at a time. However, the extent to which the value of R^2_k is increased by the addition of several symptoms may be used to determine the diagnostic value of symptoms chosen several at a time.

Thus, if additional symptoms S_q and S_r are considered, their diagnostic value with respect to the k'th disease is

$$\Delta R^2_{kqr} = R^2_{kqr} - R^2_k \qquad (6.20)$$

where

$$R^2_{kqr} = \sum_{i}^{M+2} \sum_{j}^{M+2} \bar{s}_{ki} . d_{ij} . \bar{s}_{kj} \quad . \qquad (6.21)$$

Unfortunately the number of combinations of even two symptoms often renders multiple evaluation of expressions such as (6.21) prohibitive. For this reason, for the applications discussed in this chapter it is assumed that the symptoms are selected one at a time.

6.4 Results Using Sequential Diagnosis

Using the data supplied by Scheinok, the 223 previous patients were divided into two sets. One set formed the sample to be diagnosed and was composed of the 25 unique symptom vectors discussed in Chapter 3. The other set formed the data base of previous patients needed for determination of R^2_k etc. All previous patients having symptom vectors equal to those in the sample were then removed from the data base. This ensured that all 25 symptom vectors were new patients and left 162 previous patients in the data base.

The disease of every new patient was known a priori. It was therefore decided to presume an environment in which the doctor had correctly diagnosed

the disease, and now required confirmation of that diagnosis.

The symptoms used in this application are yes/no type answers to questions. It was therefore presumed that all costs e_q were equal to unity.

6.4.1 Symptom Sequences

The diagnostic value of a symptom, as defined by (6.10), is independent of the symptoms exhibited by the new patients. Thus, using only the previous patients' records, (6.10) was used to determine a sequence of symptoms which would lead to the correct diagnosis of all previous patients having and not having each disease D_k. Such a sequence is said to be "k'th disease conscious". Additionally a non-disease-conscious sequence of symptoms was determined using ΔR^2_{Kq} of form (6.19 ; L = K).

It was assumed that the initial diagnosis would be made using the 6 symptoms $S_6,\ldots,S_{11}$. The sequence in which the remaining symptoms $(S_1,\ldots,S_5)$ are chosen is shown in Table 6.1. Note that every k'th-disease-conscious sequence is different, and different again from the non-disease-conscious sequence.

If all costs e_q are equal, each additional symptom in a k'th-disease-conscious sequence is chosen to

Disease D_k	Symptom Sequence	Number of Symptoms					
		6	7	8	9	10	11
Hiatal Hernia	D-C	S_6	S_3	S_2	S_5	S_1	S_4
Duodenal Ulcer	D-C	S_6	S_5	S_1	S_2	S_3	S_4
Gastric Ulcer	D-C	S_6	S_3	S_2	S_1	S_5	S_4
Cancer	D-C	S_6	S_5	S_1	S_3	S_2	S_4
Gallstones	D-C	S_6	S_3	S_2	S_4	S_5	S_1
Functional Disease	D-C	S_6	S_3	S_1	S_5	S_2	S_4
	N-D-C	S_6	S_3	S_5	S_1	S_2	S_4

Table 6.1 Disease-Conscious and Non-Disease-Conscious Symptom Sequences.

give the greatest improvement in the accuracy of diagnosis of patients with respect to the disease D_k. Hence for any patient p having the disease D_k, it is to be expected that such a sequence will be near optimal with respect to the number of symptoms needed to make the definitive diagnosis $D(p) = D_k$. The results of the next section support this assertion.

However, for any patient p not having the disease D_k, one of the K-1 i'th-disease-conscious sequences of additional symptoms ($i \neq k$) will lead to the definitive diagnosis $D(p) = D_i$, and therefore $D(p) = D_{\bar{k}}$. Therefore a k'th-disease-conscious sequence of additional symptoms may not be near optimal with respect to the number of symptoms needed to make the definitive diagnosis $D(p) = D_{\bar{k}}$. Fortunately the special case in which $K = 2$ is likely to be the environment of a screening clinic.

6.4.2 The Diagnosis of New Patients

The diagnosis of new patients was made using the extended disease probabilities (4.44). It was assumed that the frequency distributions formed by the sets

$$\{Z_k(n);\ D(n) = D_i\}\ ,\quad \text{all } i \leq K$$

were normal. Linear disease-symptom functions were used.

In the previous chapter it was shown that for this data the assumption of normality is false. A histogram approach was then used and satisfactory results were obtained.

In sequential diagnosis the values of the disease-symptom functions $Z_k(n)$ change whenever an additional symptom is included. Hence, if a histogram approach is used, the intervals that divide up the disease-symptom function space must also be changed. It was therefore decided that a histogram approach was too time consuming to be applicable in a sequential diagnosis environment.

The reason for not using further extended disease probabilities (4.51) was that, in the previous chapter (Table 4.4), the non-normality of the data resulted in a lower accuracy of diagnosis when including the additional probabilities $P(Z_i(n)|D_k)$, all $i \neq k$.

For simplification of computation, all parameters β_k were set to zero. The results, therefore, must be treated as indicative of what might be obtained under conditions of normality, using suitable β_k, and further extended disease probabilities.

Table 6.2 shows the values of the probabilities $P(D_k|Z_k(n^*))$, all $k \leq K$, for three new patients using

New Patient	Number of symptoms	k = 1	k = 2	k = 3	k = 4	k = 5	k = 6
1745 (k=2) duodenal ulcer	6	.0282	.6471	.3859	.0551	.0000	.1057
	11	.0217	.8573	.2432	.0130	.0000	.0316
811 (k=5) gall-stones	6	.0524	.0000	.0950	.9328	.5711	.0808
	11	.0717	.0000	.0222	.7891	.8585	.0174
560 (k=6) functional disease	6	.0277	.6588	.0916	.0000	.0792	.1599
	11	.0215	.1923	.2282	.0000	.0026	.4570

Table 6.2 Values of $P(D_k | Z_k(n^*))$ for Three New Patients Using Six and 11 Symptoms.

at first six and then 11 symptoms. The patient numbers 1745, 811 and 560 are the decimal equivalent of the binary-valued symptom vector $(S_1,\ldots,S_{11})$ which each patient exhibits. The table shows that as the number of symptoms used is increased (from 6 to 11) then, for these new patients, the correct diagnoses become more certain and the incorrect diagnoses become less certain. The table also shows that, for this data, insufficient symptoms are available to make a definitive diagnosis.

The change in the disease probabilities, as observed in the three new patients of Table 6.2, was not observed in all 25 new patients. However, there was a definite trend towards an improvement in the accuracy of diagnosis as the number of symptoms used was increased. Using six symptoms, 12 new patients were correctly diagnosed. Using 11 symptoms, 16 new patients were correctly diagnosed. The correct diagnosis was based upon the most probable disease determined.

Table 6.3 shows the value of the probability $P(D_k|Z_k(n^*))$ averaged over four new patients having duodenal ulcer ($k = 2$), four new patients having gallstones ($k = 5$) and three new patients having functional disease ($k = 6$). Average values were used in order to obtain a smoothing of the results. The average values of the disease probabilities $P(D_k|Z_k(n^*))$ for $k = 2,5,6$, are shown as each additional symptom is

Disease D_k	Symptom Sequence	Number of Symptoms					
		6	7	8	9	10	11
Duodenal Ulcer	D-C	.7576	.8838	.8852	.8903	.8897	.8916
	N-D-C	.7576	.7610	.8851	.8860	.8897	.8916
Gall-stones	D-C	.5257	.8092	.9184	.9047	.9043	.9025
	N-D-C	.5257	.8092	.8060	.8090	.9190	.9025
Functional Disease	D-C	.2004	.2574	.3274	.3689	.3630	.3615
	N-D-C	.2004	.2574	.3032	.3689	.3630	.3615

Table 6.3 Average Values of $P(D_k|Z_k(n^*))$ for Sequential Diagnosis of New Patients Known to Have D_k.

included in the appropriate k'th-disease-conscious and non-disease-conscious sequences shown in Table 6.1.

Table 6.4, prepared in part from Table 6.3, shows that the disease probabilities $P(D_k|Z_k(n^*))$ have, in nine of 13 possible instances, a greater average value when using appropriate k'th-disease-conscious sequences of additional symptoms than when using non-disease-conscious sequences. This implies (at the 83% significance level) that an appropriate k'th-disease-conscious sequence of additional symptoms can confirm a diagnosis using fewer symptoms than can a non-disease-conscious sequence.

The average values of $P(D_k|Z_k(n^*))$ for each disease D_k were determined using 18 of the 25 new patients. The remaining seven new patients in the sample were not included in determining the average values of the disease probabilities since their disease probabilities changed inconsistently as the number of symptoms used was increased. Presumably this could be explained by their location in the disease-symptom space relative to the previous patients, and that in consequence the hyperplanes, defined by the linear disease-symptom functions, were not suitably oriented with respect to these new patients. Also all of these seven new patients were incorrectly diagnosed even when using all 11 symptoms $(S_1,....,S_{11})$.

	Number of Symptoms					
Disease D_K	6	7	8	9	10	11
Hiatal Hernia	.	.	-0.0009	-0.0015	.	.
Duodenal Ulcer	.	+0.1228	+0.0001	+0.0043	.	.
Gastric Ulcer	.	.	+0.0115	+0.0017	.	.
Cancer	.	-0.0147	+0.0111	.	.	.
Gallstones	.	.	+0.1124	+0.0957	-0.0147	.
Functional Disease	.	.	+0.0242	.	.	.

Table 6.4 Probability Difference Table.

6.4.3 An On-line Interactive System

The preceding application of sequential diagnosis assumed that the doctor had correctly diagnosed the disease, and now required confirmation of that diagnosis. Thus it was possible to use a predetermined k'th-disease-conscious sequence of additional symptoms.

In general the doctor will not make such a definitive diagnosis. Rather he will go through a procedure of considering several diseases as possibilities, eliminating some and including others as the values of the additional symptoms become known. An example of such a procedure, as it might appear on a computer terminal in the doctor's office, is presented in Figure 6.1.

The figure was derived from the sequential diagnosis of a new patient, the decimal equivalent of whose binary valued symptom vector is 1745 (see Table 6.2). Since, when using this data, it is not possible to make a definitive diagnosis, no disease was considered significant unless the probability exceeded 0.1. Such a figure might not, of course, be used in actual application.

At "A" the doctor enters which classes of disease are to be considered for diagnosis. The appropriate disease-symptom matrices $\bar{S}$, G_k and $\bar{S}_k$ are

```
WHAT CLASSES OF DISEASE DO YOU WISH TO CONSIDER FOR
DIAGNOSIS?

          upper abdominal pain                    (A)
          #

WHICH SYMPTOMS WILL BE USED IN ORDER TO MAKE THE FIRST
DIAGNOSIS?

          brief irregular
          food relief
          food aggravation
          position aggravation                    (B)
          weight loss
          persistence
          #

IN THE SAME ORDER AS ABOVE, ENTER THE VALUES OF THESE
SYMPTOMS AS OBSERVED ON THE PATIENT.

          absent
          present
          absent                                  (C)
          absent
          absent
          present

THE MOST SIGNIFICANT DISEASES DIAGNOSED, TOGETHER WITH
THEIR 'PROBABILITIES' ARE

          DUODENAL ULCER      0.6471
          GASTRIC ULCER       0.3859              (D)
          FUNCTIONAL DISEASE  0.1057

WHICH DISEASES DO YOU WISH TO CONSIDER FOR FURTHER
DIAGNOSIS?

          duodenal ulcer
          cancer                                  (E)
          #

THE SYMPTOM MOST APPROPRIATE FOR DIAGNOSIS OF THESE
DISEASES ARE

          CLUSTERS
     OR, SEX                                      (F)

WHICH SYMPTOM WILL YOU USE?

          clusters                                (G)
```

Figure 6.1 Sequential Diagnosis of A New Patient

```
ENTER THE VALUE OF THIS SYMPTOM AS OBSERVED ON THE PATIENT.

                present

THE MOST SIGNIFICANT DISEASES DIAGNOSED, TOGETHER WITH
THEIR 'PROBABILITIES' ARE

                DUODENAL ULCER    0.7885
                GASTRIC ULCER     0.2876

WHICH DISEASES DO YOU WISH TO CONSIDER FOR FURTHER
DIAGNOSIS?

                duodenal ulcer
                #

THE SYMPTOM MOST APPROPRIATE FOR DIAGNOSIS OF THIS
DISEASE IS

                SEX
          OR,   EPIGASTRIC PAIN

WHICH SYMPTOM WILL YOU USE?

                sex

ENTER THE VALUE OF THIS SYMPTOM AS OBSERVED ON THE PATIENT.

                male

THE MOST SIGNIFICANT DISEASES DIAGNOSED, TOGETHER WITH
THEIR 'PROBABILITIES' ARE

                DUODENAL ULCER    0.8460
                GASTRIC ULCER     0.2829

WHICH DISEASES DO YOU WISH TO CONSIDER FOR FURTHER
DIAGNOSIS?

                none                                    (H)
```

Figure 6.1 (cont'd)

then automatically loaded into the computer memory. At "B" the doctor declares which symptoms are to be used in making the initial diagnosis. The coefficients C_{km} for these symptoms are then computed using suitable values of β_k.

Using the values of the symptoms, as entered at "C", the computer determines which of all the diseases considered for diagnosis are significant, point "D". The doctor then decides which diseases to consider when selecting the next symptom. He may, of course, include additional diseases if he wishes, at point "E".

By comparing the relative values of ΔR^2_{Lq} for each remaining symptom, the computer displays, point "F", in order of diagnostic value, the additional symptoms most appropriate for the diagnosis of the diseases entered at "E". When the doctor has chosen one of these symptoms or another symptom not listed, point "G", the coefficients C_{km}, using suitable β_k, are redetermined for all diseases in the classes, as entered at "A".

The procedure so repeats until a definitive diagnosis is made, or the doctor decides to terminate the procedure, point "H". A listing of the computer program from which this figure was derived is presented in Appendix 4.

6.5 Conclusion

The method for sequential diagnosis developed in this chapter is extremely flexible. Sequences of symptoms can be determined that will lead to any patient p being diagnosed as having, or not having, any specified disease D_k. Further these sequences would seem to be near optimal with respect to the number of symptoms required to make the definitive diagnosis $D(p) = D_k$. Such sequences offer the advantage that prior to any new patients being diagnosed, the coefficients C_k, using optimum β_k, can be determined. Thus the day-to-day determination of disease probabilities of patients attending, say, a screening clinic, could be handled by a mini computer.

Alternatively the method is capable of choosing additional symptoms on the basis of the current diagnosis, or at the discretion of the doctor. To do so, of course, requires considerable computing power. But with the wide availability of time-sharing systems, any doctor can get access to such facilities by way of a computer terminal in his office.

CHAPTER 7

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

This thesis has concentrated on the formulation and application of three methods for automatic diagnosis of disease. These methods are all related by the concept of disease-symptom functions and the expressions used to determine the coefficients thereof.

Each disease has its own disease-symptom function, being any mathematical expression of the symptoms. The particular class of generalized linear disease-symptom functions was considered. The advantages of such functions are; they make no assumptions as to the statistical independence of the symptoms; they may be defined to allow for any non-linear and/or interactive order of dependence between symptoms and disease; the symptoms may be multivalued; and, by the addition of a constant, they are independent of the quantization of the symptoms.

In Chapter 3 patients were diagnosed as having the disease corresponding to that disease— symptom function having the largest value, as determined from the patient's symptoms. A method was formulated for determining the coefficients of the disease-symptom functions. This method introduced the concept of using

parameters, alpha and beta, to change the coefficients linearly and non-linearly in order to obtain a maximum number of correct diagnoses of patients in the data base. When applied to a data base of several hundred gastroenterological patients, each known to have one of six diseases, an accuracy of diagnosis of 80.3% of the previous patients was obtained. This compared with 74.9% using Bayes' theorem, and 70.8% using a previously formulated method employing disease-symptom functions.

A distinction was made between new and previous patients. It was empirically shown that the accuracy of diagnosis of previous symptom vectors (and therefore patients) does not directly relate to the accuracy of diagnosis of new symptom vectors. However, for this data, as the size of the data base increased, so did the accuracy of diagnosis of new symptom vectors. It was argued that this could be explained by the need for the data base to be representative of the new symptom vectors, and that as the data base grew in size, so it became more representative. The conclusion reached was that, when using automatic methods for diagnosis, the data base should be as large as possible.

A comparison of using linear and quadratic disease-symptom functions revealed that the correct diagnosis of all previous patients does not necessarily

lead to the correct diagnosis of all new patients. The problem of how to determine the order of dependence between symptoms and disease that will result in the correct diagnosis of a maximum number of new patients was not discussed. This is one area justifying further research.

In Chapter 4 disease probabilities were determined by applying Bayes' theorem to the values of the disease-symptom functions, thereby retaining all the advantages of disease-symptom functions. However, this approach assumes that, for each disease, the disease-symptom functions are independent. But, it was argued, this assumption is more plausible than the conventional assumption that it is the symptoms which are independent.

Consideration was given to the problem of determining the coefficients of the disease-symptom functions in order that the disease probabilities result in the correct diagnosis of a maximum number of patients in the data base. No general solution to this problem is known. A constrained solution was found, by using each parameter beta, to maximize an expression representative of the requirement of maximizing the number of previous patients correctly diagnosed. By assuming that the frequency distributions of the values of the disease-symptom functions for patients in

the data base are normal, estimates for the value of each parameter beta were determined. A directed search strategy for finding the optimum value of each beta was then proposed. An unconstrained solution to this problem would result in a further improvement in the accuracy of diagnosis.

The method was used to determine the most probable disease of each patient in the data base of gastroenterological patients. The assumption of normality was shown to be false, and the disease probabilities were determined using a histogram approach. It was shown that the optimum values of each parameter beta, as found using the assumption of normality, raised the accuracy of diagnosis from 79.4% to 82.5%. The results of this limited application suggest that, for each disease, the disease-symptom functions may be assumed to be independent, and that the method of determining the optimum value of each beta is sound. In the absence of further applications, this is the strongest conclusion that can be made.

In Chapter 5 consideration was given to the manner in which each method, for automatic diagnosis, divides up the disease-symptom space into regions for classification. Reasons were given as to why higher accuracies, of diagnosis of previous patients, were obtained when using the method formulated in Chapter 3

(80.3%) and Chapter 4 (82.5 with when using other methods (70.8% and 74.

A method for sequenti is was developed in Chapter 6. Additional sy e chosen according to their diagnostic valu cost. The diagnostic value of each sym efined as measuring the increase in ac diagnosis of previous patients, when usin probabilities, that results from the additi symptom. It was shown that, with some si n, such a measure could be obtained by conside pression used to determine the coefficients o se-symptom functions. This result implied cy between disease-symptom functions, d abilities (assuming normality), and di lue. Further, all the advantages of diseas unctions were retained.

The diagnostic value, , was said to be disease conscious, being th respect to each disease. A disease-con nce of additional symptoms was then determined disease in the data base of gastroenterolog s. It was argued, and empirical result the assertion, that if all costs were equal onscious sequence of additional sympt near optimal with respect to the number o eeded to make

a definitive diagnosis of any patient having that disease.

The method for sequential diagnosis was also shown to have potential as an on-line interactive system; the doctor and the computer communicating with regard to the selection of additional symptoms, current disease probabilities, etc. There is, however, considerable overhead involved in matrix inversion, and further investigation is needed in order to determine the feasibility of such a system.

Another area deserving investigation follows from the distinction made between new and previous patients. Actual application of automatic diagnosis requires that when presented with a patient to be diagnosed the patient's disease be determined from the patient's symptom vector. Hence, if the data base, of all the previous patients' records, is ordered using symptom vectors as a key, then it may be feasible to search the data base to determine whether or not the patient to be diagnosed exhibits a symptom vector equal to that of any previous patients. For if this is so, the disease of the patient to be diagnosed is known. If no such previous patients can be found, then the patient to be diagnosed is truly a new patient. Only then need the methods, proposed in this thesis, be used.

From the point of view of using the decision rule in Chapter 3, and the assumption made in Chapter 4, that the diseases are mutually exclusive, the methods developed in this thesis require that every patient has only one disease. It should be possible to relax this requirement. But to do so will require appropriate redefinition of the decision rule used in Chapter 3, and reformulation of most of Chapter 4, together with an appropriate interpretation of the disease probabilities. Such an undertaking must also allow for the fact that the manner in which symptoms are exhibited in patients having several diseases may be very different from that exhibited by patients having one disease. However, the requirement that each patient have only one disease is not overly restrictive, since any patient having several diseases may be regarded as having one new disease.

Methods for automatic diagnosis of disease have potential for application in two areas of medicine. The first concerns the training of medical students. Studies in this area have been performed by Harless (1971), using a model called "Case" (Computer-Aided Simulation of the Clinical Encounter), and by Schneiderman (1972), using the "Diagnosis Game". In these studies the student attempts to diagnose a "patient" by

"observation" of symptoms displayed on a computer terminal. Audio-visual equipment is often used to simulate the true doctor/patient environment.

The second area concerns the use of computers as an aid to the doctor. Automatic diagnosis has the convenience that the interview and administering of tests (perhaps in response to the computer's requests), can be performed by paramedical personnel. The doctor can interview the patient at the end of the computer diatnosis i.e. when the test results and disease probabilities are known. The doctor is free to overrule the computer diagnosis if he wishes, or commence sequential diagnosis if necessary, and the list of diseases, in order of probability, always reminds the doctor of all possible diseases. Under such an arrangement each patient has the security of a diagnosis made by a doctor, guided by a computer having access to large volumes of data.

BIBLIOGRAPHY

Anderson, T.W. (1958) An Introduction To Multivariate Statistical Analysis, Wiley, p.134.

Bahadur, R.R. (1961) 'A representation of the joint distribution of responses to n dichotomous items'. In: Studies In Item Analysis And Prediction, Stanford University Press, p.158.

Bailey, N.T.J. (1965) 'Probability methods of diagnosis based on small samples'. In: Mathematics And Computer Science In Biology And Medicine, Her Majesty's Stationary Office, p.103.

Best, W.R. 'Allowance for non-independence of symptoms in medical diagnosis'. Submitted to Biometrics.

Boyle, J.A., Greig, W.R., Franklin, D.A., Harden, R.McG., Buchanan, W.W., and McGirr, E.M. (1966) 'Construction of a model for computer assisted diagnosis: Application to the problem of non-toxic goitre', Quarterly Journal of Medicine, New Series, 35, p.565.

Carl, J.W., and Hall, C.F. (1972) 'The application of filtered transforms to the general classification problem', I.E.E.E. Transactions on Computers, C-21, p.785.

Clendening, L., and Hashinger, E.H. (1947) Methods of Diagnosis, Mosby, p.59.

Croft, D.J. (1972) 'Is computerized diagnosis possible?', Computers and Biomedical Research, 5, p.351.

Crooks, J., Murray, I.P.C., and Wayne, E.J. (1959) 'Statistical methods applied to the clinical diagnosis of thyrotoxicosis', Quarterly Journal Medicine, New Series, 38, p.211.

Cumberbatch, J., and Heaps, H.S. (1973) 'Application of a Non-Bayesian approach to computer aided diagnosis for upper abdominal pain', Bio-Medical Computing, 4, p.105.

Cumberbatch, J., Leung, K.V., and Heaps, H.S. (1974) 'A non-probabilistic method for automatic medical diagnosis', Bio-Medical Computing, 5, p.133.

Davies, P. (1972) 'Symptom diagnosis using Bahadur's distribution', Bio-Medical Computing, 3, p.307.

Duda, R.O., and Hart, P.E. (1973) Pattern Classification And Scene Analysis, Wiley.

Dwork, B.M. (1950) 'Detection of a pulse superimposed on fluctuation noise', Proceedings of I.R.E., p.771.

Feller, W. (1966) An Introduction To Probability Theory, Vol. 2, Wiley, p.85.

Fisher, R.A. (1936) 'The use of multiple measurements in taxanomic problems', Annals of Eugenics, 7, p.179.

Freeman, F.R. (1972) 'Medical diagnosis: comparison of human and computer logic', Bio-Medical Computing, 3, p.217.

Gleser, M.A., and Collen, M.F. (1972) 'Towards automated medical decisions', Computers and Biomedical Research, 5, p.180.

Gorry, G.A., and Barnet, G.O. (1968) 'Experiences with a model of sequential diagnosis', Computers and Biomedical Research, 1, p.490.

Greenhouse, S.W. (1954) 'On the problem of discrimination between statistical populations', M.A. Thesis, George Washington University, See Kullback (1968, p.205).

Harless, W.G., Drennon, G.G., Marxer, J.J., Root, J.A., and Miller, G. (1971) 'Case: a computer-aided simulation of the clinical encounter', Journal of Medical Education, 46, p.443.

Heaps, H.S. (1973) 'A general theory for automatic diagnosis', Kybernetes, 2, p.3.

Hughes, G.F. (1968) 'On the mean accuracy of statistical pattern recognizers', I.E.E.E. Transactions on Information Theory, IT-14, p.55.

Jeffreys, H. (1948) Theory of Probability, Oxford University Press, p.158.

Kullback, S. (1968) Information Theory and Statistics, Wiley, p.6.

Ledley, R.S., and Lusted, L.D. (1959) 'Reasoning foundations of medical diagnosis', Science, 130, No. 3366, p.9.

Ledley, R.S., and Lusted, L.B. (1960) 'The use of electronic computers in medical data processing: Aids in diagnosis, current information retrieval, and medical record keeping', I.R.E. Transactions on Medical Electronics, 7, p.31.

Levinson, N. (1947) 'The Wiener R.M.S. Error criterion in filter design and prediction', Journal of Mathematics and Physics, 15, p.261.

Marascuilo, L.A. (1971) Statistical Methods For Behavioral Science Research, McGraw-Hill, p.258.

Nilsson, N.J. (1965) Learning Machines, McGraw-Hill, p.15.

Pipberger, H.V., Klingeman, J.D., and Cosma, J. (1968) 'Computer evaluation of statistical properties of clinical information in the differential diagnosis of chest pain', Methods for Information in Medicine, 7, p.79.

Reale, A., Maccaro, G.A., Rocca, E., D'Intino, S., Gioffre, P., Vestri, A., and Motelese, M. (1968) 'Computer diagnosis of congenital heart disease', Computers and Biomedical Research, 1, p.533.

Rinaldo, J.A., Scheinok, P., Rupe, C.E. (1963) 'Symptom Diagnosis: A mathematical analysis of epigastric pain', Annals of Internal Medicine, 59, p.145.

Royce, J.I. (1973) Multivariate Analysis And Psychological Theory, Academic Press.

Savage, L.J. (1954) The Foundations of Statistics, Wiley, p.50.

Scheinok, P.A., and Rinaldo, J.A. (1967) 'Symptom diagnosis: Optimal subsets for upper abdominal pain', Computers and Biomedical Research, 1, p. 221.

Scheinok, P.A., and Rinaldo, J.A. (1968) 'Symptom diagnosis: A comparison of mathematical models related to upper abdominal pain', Computers and Biomedical Research, 1, p.457.

Scheinok, P.A. (1969) 'Symptom diagnosis, dependence, independence and entropy', Materia Medica Polonia, FASC.3/4, p.14.

Scheinok, P.A. (1972a) 'Symptom diagnosis: Bayes' theorem and Bahadur's distribution', Bio-Medical Computing, 3, p.17.

Scheinok, P.A. (1972b) Personal communication to Prof. H.S. Heaps.

Schneiderman, H., and Muller, R.L. (1972) 'The diagnosis game', Journal of the American Medical Association, 219, p.333.

Sebestyen, G.S. (1962) Decision Making Processes In Pattern Recognition, Macmillan, p.40.

Taylor, T.R. (1970) 'Computers in medicine in the decade of the 1970's', Scottish Medical Journal, 15, p.353.

Taylor, T.R., Shields, S., and Black, R. (1972) 'Study of cost-conscious computer assisted diagnosis in thyroid disease', The Lancet, July 8, 1972, p.79.

Van der Geer, J.P. (1971) Introduction To Multivariate Analysis For The Social Sciences, Freemon, p.260.

Vanderplas, J.M. (1967) 'A method for determining probabilities for correct use of Bayes' theorem in medical diagnosis', Computers and Biomedical Research, 1, p.215.

Wang, Y. (1972) 'Computers in medical diagnosis', Critical Reviews In Radiological Sciences, The Chemical Rubber Company, April 1972, p.200.

Warner, H.R., Toronto, A.F., Veasey, G., and Stephenson, R. (1961) 'A mathematical approach to medical diagnosis', Journal of the American Medical Association, 177, No. 3, p.177.

Wilson, K.V. (1973) 'Linear regression equations as behaviour models', see Royce, J.I. (1973, p.45).

APPENDIX 1

First note that

$$R_k = \frac{\bar{z}_{k1}}{[\frac{1}{N}\sum_n z_k^2(n)]^{\frac{1}{2}}} = \frac{C_k^T \bar{S}_k}{[C_k^T \bar{S} C_k]^{\frac{1}{2}}} \tag{A1.1}$$

(by (4.12), (4.13), (4.14), (4.15) and (4.31)) and that

$$r_k = \frac{\sigma_{k1}}{\bar{z}_{k1}} = \frac{[C_k^T G_k C_k]^{\frac{1}{2}}}{C_k^T \bar{S}_k} \tag{A1.2}$$

(by (4.12) and (4.14)). Hence,

$$Q_k = R_k - \beta_k' r_k \tag{A1.3}$$

may be written in the matrix and vector form

$$Q_k = \frac{C_k^T \bar{S}_k}{[C_k^T \bar{S} C_k]^{\frac{1}{2}}} - \beta_k' \frac{[C_k^T G_k C_k]^{\frac{1}{2}}}{C_k^T \bar{S}_k} \quad . \tag{A1.4}$$

In order to determine the derivative of Q_k with respect to C_k the three relations

$$\frac{d}{dC_k}(C_k^T \bar{S}_k) = \bar{S}_k \quad , \tag{A1.5}$$

$$\frac{d}{dC_k}(C_k^T \bar{S} C_k) = 2\bar{S}C_k \quad , \tag{A1.6}$$

and

$$\frac{d}{dC_k}(C_k^T G_k C_k) = 2G_k C_k \tag{A1.7}$$

will be used. The relations (A1.6) and (A1.7) are due

to Anderson (1958, p.347). Accordingly

$$\frac{dQ_k}{dC_k} = [C_k^T \bar{S} k].[-\tfrac{1}{2}][C_k^T \bar{S} C_k]^{-\frac{3}{2}}[2\bar{S}C_k] + \bar{S}_k[C_k^T \bar{S} C_k]^{-\frac{1}{2}}$$

$$- \beta_k'[C_k^T G_k C_k]^{\frac{1}{2}}[-1][C_k^T \bar{S}_k]^{-2}\ \bar{S}_k$$

$$- \beta_k'[\tfrac{1}{2}][C_k^T G_k C_k]^{-\frac{1}{2}}[2G_k C_k][C_k^T \bar{S}_k]^{-1} \quad . \qquad \text{(A1.8)}$$

It follows that (A1.8) is stationary when

$$\left(\bar{S} + \beta_k' \frac{[C_k^T \bar{S} C_k]^{\frac{3}{2}} G_k}{[C_k^T G_k C_k]^{\frac{1}{2}}[C_k^T \bar{S}_k]^2}\right) C_k =$$

$$\frac{[C_k^T \bar{S} C_k]^{\frac{1}{2}}}{C_k^T \bar{S}_k}\left(\frac{1}{[C_k^T \bar{S} C_k]^{\frac{1}{2}}} + \beta_k' \frac{[C_k^T G_k C_k]^{\frac{1}{2}}}{[C_k^T \bar{S}_k]^2}\right)\bar{S}_k \quad . \qquad \text{(A1.9)}$$

Since

$$r_k R_k^3 = \frac{[C_k^T G_k C_k]^{\frac{1}{2}}[C_k^T \bar{S}_k]^2}{[C_k^T \bar{S} C_k]^{\frac{3}{2}}} \qquad \text{(A1.10)}$$

the required coefficients are given by

$$C_k = \alpha_k[\bar{S} + \beta_k G_k]^{-1}\ \bar{S}_k \qquad \text{(A1.11)}$$

where α_k is scalar and

$$\beta_k = \frac{\beta_k'}{r_k R_k^3} \quad . \qquad \text{(A1.12)}$$

APPENDIX 2

Equation (4.62) may be written in matrix and vector form as

$$J_k' = \frac{[C_k^T(\bar{S}_k - \bar{S}_{\bar{k}})]^2}{2C_k^T G_k C_k} + \frac{[C_k^T(\bar{S}_k - \bar{S}_{\bar{k}})]^2}{2C_k^T G_{\bar{k}} C_k} \qquad \text{(A2.1)}$$

(by (4.12), (4.13), (4.14) and 4.15)). Putting

$$D = \bar{S}_k - \bar{S}_{\bar{k}} \qquad \text{(A2.2)}$$

it follows that

$$J_k' = \frac{C_k^T D\, D^T C_k}{2C_k^T G_k C_k} + \frac{C_k^T D\, D^T C_k}{2C_k^T G_{\bar{k}} C_k} \quad . \qquad \text{(A2.3)}$$

Using the relation

$$\frac{d}{dC_k}[C_k^T X C_k] = 2X C_k \qquad \text{(A2.4)}$$

where X is any symmetric M×M matrix (Anderson, 1958, p.347), the derivative of J_k' (A2.3) with respect to C_k is given by

$$\frac{dJ_k'}{dC_k} = \frac{C_k^T G_k C_k DD^T C_k - C_k^T DD^T C_k G_k C_k}{[C_k^T G_k C_k]^2}$$

$$+ \frac{C_k^T G_{\bar{k}} C_k DD^T C_k - C_k^T DD^T C_k G_{\bar{k}} C_k}{[C_k^T G_{\bar{k}} C_k]^2} \qquad \text{(A2.5)}$$

Putting

$$a_k = C_k^T G_k C_k \quad , \qquad \text{(A2.6)}$$

$$a_{\bar{k}} = C_k^T G_{\bar{k}} C_k \qquad \text{(A2.7)}$$

and

$$b = D^T C_k = C_k^T D \qquad \text{(A2.8)}$$

it follows that (A2.5) is stationary when

$$\frac{a_k D b - b^2 G_k C_k}{a_k^2} + \frac{a_{\bar{k}} D\, b - b^2 G_{\bar{k}} C_k}{a_{\bar{k}}^2} \qquad \text{(A2.9)}$$

which is satisfied when

$$(a_{\bar{k}}^2 a_k + a_{\bar{k}} a_k^2) D = b(a_{\bar{k}}^2 G_k + a_k^2 G_{\bar{k}}) C_k \ . \qquad \text{(A.10)}$$

Hence

$$C_k = a_4 (b_4 G_k + G_{\bar{k}})^{-1} D \qquad \text{(A2.11)}$$

where

$$a_4 = \frac{a_{\bar{k}}^2 + a_{\bar{k}} a_k}{b\, a_k} \ , \qquad \text{(A2.12)}$$

which is scalar, and

$$b_4 = \frac{a_{\bar{k}}^2}{a_k^2} = \left[\frac{C_k^T G_{\bar{k}} C_k}{C_k^T G_k C_k}\right]^2 \ . \qquad \text{(A2.13)}$$

The expression J_k' (A2.3) is the sum of two generalized Rayleigh quotients. Since a Rayleigh quotient is a concave downward function having one maximum value, and there is only one value of the C_k (A2.11) for which the first derivative of J_k' is zero, it follows that (A2.11) are the coefficients which maximize J_k'.

APPENDIX 3

Equation (4.62) is of the form

$$J_k' = \frac{(\bar{z}_{k1} - \bar{z}_{k2})^2}{2\sigma_{k1}^2} + \frac{(\bar{z}_{k1} - \bar{z}_{k2})^2}{2\sigma_{k2}^2} \qquad \text{(A3.1)}$$

In the instance that $\beta_k = 0$ and it is assumed that $\sigma_{k1}^2 = \sigma_{k2}^2 = C_k^T G C_k$, where $G = (N_{k1}G_k + N_{k2}G_{\bar{k}})/N$ then (A3.1) may be written in matrix and vector form as

$$J_k' = \frac{[C_k^T(\bar{S}_k - \bar{S}_{\bar{k}})]^2}{C_k^T G C_k} \quad . \qquad \text{(A3.2)}$$

For simplicity of notation let the suffices k1 and k be denoted by 1, and the suffices k2 and $\bar{k}$ be denoted by 2. Then denoting (A3.2) by J;

$$J = \frac{[C_1^T(\bar{S}_1 - \bar{S}_2)]^2}{C_1^T G C_1} \qquad \text{(A3.3)}$$

so that since J is independent of the constant C_{10} (4.34) and

$$C_1 = \frac{\alpha_1 X N_2}{N} G^{-1} (\bar{S}_1 - \bar{S}_2) \qquad \text{(A3.4)}$$

(by (4.38)), substitution into (A3.3) shows that

$$J = \frac{[(\bar{S}_1-\bar{S}_2)^T G^{-1}(\bar{S}_1-\bar{S}_2)][C_1^T(\bar{S}_1-\bar{S}_2)]}{(\bar{S}_1-\bar{S}_2)^T G^{-1} G G^{-1}(\bar{S}_1-\bar{S}_2)\ \frac{\alpha_1 X N_2}{N}}$$

$$= \frac{C_1^T(\bar{S}_1-\bar{S}_2)}{[\alpha_1 X N_2/N]} = \frac{N(\bar{Z}_1-\bar{Z}_2)}{\alpha_1 X N_2} \qquad \text{(A3.5)}$$

(by 4.12) and (4.14)). Also, since

$$G = \frac{N_1 G_1 + N_2 G_2}{N} \tag{A3.6}$$

(by (4.60)) it follows that

$$J = \frac{N(\bar{Z}_1 - \bar{Z}_2)^2}{N_1\sigma_1^2 + N_2\sigma_2^2} \quad . \tag{A3.7}$$

Accordingly (A3.5) and (A3.7) may be used to derive the relation

$$N_1\sigma_1^2 + N_2\sigma_2^2 = \alpha_1 X N_2(\bar{Z}_1 - \bar{Z}_2) \quad . \tag{A3.8}$$

If $\bar{S}_1 \neq \bar{S}_2$ then (4.37) shows that

$$\frac{N_1}{N}(\bar{Z}_1 - \bar{Z}_2) = \alpha_1(1 - X) \tag{A3.9}$$

so that

$$X = \frac{N\alpha_1 - N_1(\bar{Z}_1 - \bar{Z}_2)}{N\alpha_1} \quad . \tag{A3.10}$$

Denoting R_k^2 ($\beta_k = 0$) by R^2 it may similarly be shown that

$$R^2 = \frac{N\bar{Z}_1^2}{[N_1\sigma_1^2 + N_2\sigma_2^2 + N_1\bar{Z}_1^2 + N_2\bar{Z}_2^2]} = \frac{\bar{Z}_1}{\alpha_1} \tag{A3.11}$$

(by (6.2) and (3.28)). Thus, (A3.11) provides the relation

$$N_1\sigma_1^2 + N_2\sigma_2^2 = \alpha_1 N\bar{Z}_1 - N_1\bar{Z}_1^2 - N_2\bar{Z}_2^2 \quad . \tag{A3.12}$$

Equations (A3.8), (A3.9) and (A3.10) may be solved to give

$$(N_1\bar{z}_1 + N_2\bar{z}_2)^2 = \alpha_1(N_1 + N_2)(N_1\bar{z}_1 + N_2\bar{z}_2) \ . \qquad (A3.13)$$

If $N_1\bar{z}_1 + N_2\bar{z}_2 = 0$ then $J < 0$, which is false. Hence, since $N_1 + N_2 = N$, it follows that

$$N_1\bar{z}_1 + N_2\bar{z}_2 = \alpha_1 N \ , \qquad (A3.14)$$

and therefore

$$\bar{z}_1 - \bar{z}_2 = \frac{N}{N_2}(\bar{z}_1 - \alpha_1) \ . \qquad (A3.15)$$

Substituting (A3.10) into (A3.5) and using (A3.15) gives

$$J = \frac{N^2(\bar{z}_1 - \alpha_1)}{N_2(N_2\alpha_1 - N_1(\bar{z}_1 - \alpha_1))} \qquad (A3.16)$$

so that since $\bar{z}_1 = \alpha_1 R^2$ (by (A3.11)), and with some rearrangement

$$R^2 = 1 + \frac{N_2^2}{N_1N_2 + \frac{N^2}{J}} \ . \qquad (A3.17)$$

APPENDIX 4

```
C
C
C              SIGNAL TO SIGNAL PLUS NOISE RATIO TECHNIQUE
C
C         CALCULATING:    LINEAR DISEASE-SYMPTOM FUNCTIONS,
C                         EXTENDED DISEASE PROBABILITIES, AND
C                         FURTHER EXTENDED DISEASE PROBABILITIES
C                         FOR PREVIOUS AND NEW PATIENTS.
C
C       INITIALIZATION PARAMATERS, ARRAYS, ETC.
      DIMENSION  D(6,300), SIJ(12,12), SIK(12,6), AIJK(12,12,6),WMSR(6)
      DIMENSION  Q(12),R(12), P(19),E(18), F(15), QM(12), Y(6), ENT(12)
      DIMENSION ACC(6), G(12,12), QN(6,6), RN(6,6), PAVG(18), PMEAN(12)
      REAL*8 A(12,12), T(12,12), B(12,12), C(12,6), BT(6)
      INTEGER IP(24), S(14,300), DT(6,300)
      DO 01090 I=1,6
01090 F(I)=0
      DO 01150 I=1,12
      PAVG(I)=0.0
      DO 01150 J=1,12
      SIJ(I,J)=0
      DO 01150 K=1,6
      AIJK(I,J,K)=0
01150 SIK(I,K)=0
C
C       NUMBER OF DISEASES
      KD=6
C
C       READ IN DATA CARDS OF PREVIOUS PATIENTS SYMPTOMS
02030 FORMAT(12(I1,2X), 40X, I4)
      N=1
02060 READ (5, 02030, END=02090) (S(I,N),I=1,13)
      S(14,N)=S(13,N)
      S(13,N)=1
      F(S(1,N))=F(S(1,N))+1
      N=N+1
      GO TO 02060
02090 N=N-1
C
C       READ IN DATA CARDS OF NEW PATIENTS SYMPTOMS
      NP=N+1
02160 READ (5, 02030, END=02190) (S(I,NP),I=1,13)
      S(14,NP)=S(13,NP)
      S(13,NP)=1
      NP=NP+1
      GO TO 02160
02190 NP=NP-1
C
C       CREATE SIJ MATRIX
03030 DO 03070 I=1,12
      DO 03070 J=1,12
      DO 03070 K=1,N
03070 SIJ(I,J)=SIJ(I,J)+S(I+1,K)*S(J+1,K)
C
C       CREATE SIK MATRIX
```

```
      DO 04080 I=1,12
      DO 04080 K=1,KD
      DO 04070 J=1,N
04070 IF (S(1,J).EQ.K) SIK(I,K)=SIK(I,K)+S(I+1,J)
04080 SIK(I,K)=SIK(I,K)*N/F(K)
C
C     CREATE COVARIANCE -GIJK- MATRIX
      DO 05090 K=1,KD
      DO 05090 I=1,12
      DO 05090 J=1,12
      AIJK(I,J,K)=0
      DO 05080 L=1,N
05080 IF (S(1,L).EQ.K) AIJK(I,J,K)=AIJK(I,J,K)+S(I+1,L)*S(J+1,L)
05090 AIJK(I,J,K)=AIJK(I,J,K)*N/F(K)-SIK(I,K)*SIK(J,K)/N
C
C     PRINT OUT
06030 FORMAT ('1','MATRIX PRINT OUT GIJK, SIJ AND SIK')
      WRITE(6,06030)
06050 FORMAT (' ', 12(F4.0,1X))
06060 FORMAT ('+', 70X, 12(F4.0,1X))
06070 FORMAT ('+', 70X,  6(F4.0,1X))
      K=2
06090 DO 06120 I=1,12
      K=K-1
      WRITE(6,06050) (AIJK(I,J,K), J=1,12)
      K=K+1
06120 WRITE(6,06060) (AIJK(I,J,K), J=1,12)
06130 FORMAT ('      ')
06140 WRITE(6,06130)
      K=K+2
      IF (K.LT.7) GO TO 06090
      DO 06170 I=1,12
      WRITE(6,06050) (SIJ(I,J),J=1,12)
06170 WRITE(6,06070) (SIK(I,K), K=1,6)
06200 FORMAT ('1', 'COEFICIENTS FOR DISEASES')
06210 FORMAT('0  BETA   D',11X,'C1',7X,'C2',7X,'C3',7X,'C4',7X,'C5',7X,
     1        'C6',7X,'C7',7X,'C8',7X,'C9',7X,'C10',6X,'C11',6X,'C')
C
C     INITIAL NUMBER OF SYMPTOMS USED
      NOSYM=11
      M=12-NOSYM+2
      MQ=M-1
C
C     INITIAL VALUES OF PARAMATER BETA
      BTINCR=+0.2
06300 DO 06310 I=1,6
06310 BT(I)=0.0
C
C     CONTROL LOOP FOR COEFICIENTS OF EACH DISEASE
07030 WRITE(6,06200)
      WRITE (6,06210)
      K=0
07040 K=K+1
      IF (K.GT.KD)GO TO 13010
      IF (K.EQ.1) GO TO 07090
```

```
      IF (BT(K).NE.0.0) GO TO 07090
      IF (BT(K).EQ.BT(K-1)) GO TO 09030
07090 CONTINUE
C
C     CALCULATE THE A MATRIX AND INVERT IT
      DO 08050 I=1,12
      DO 08050 J=1,12
08050 A(I,J)=SIJ(I,J)+BT(K)*AIJK(I,J,K)
      IM=M-1
      DO 08140 I=1,IM
      DO 08140 J=1,12
      IF (I.NE.MQ) GO TO 08110
      IF (J.EQ.MQ) GO TO 08140
      IF (J.GE.M) GO TO 08140
08110 A(I,J)=0.0
      IF (I.EQ.J) A(I,J)=1.0
      A(J,I)=A(I,J)
08140 CONTINUE
      CALL INV(12,12,A,IP,12,T)
C
C     MULTIPLY INVERT OF A WITH A AND CHECK FOR SINGULARITY
      DO 08260 I=1,12
      DO 08260 J=1,12
      B(I,J)=0
      DO 08220 L=1,12
08220 B(I,J)=B(I,J)+A(I,L)*T(L,J)
      IF (I.EQ.J)B(I,J)=B(I,J)-1.0
      IF (B(I,J).GT.0.001) WRITE(6,08270) (I,J,K)
      IF (B(I,J).LT.-.001) WRITE(6,08270) (I,J,K)
08260 CONTINUE
08270 FORMAT ('0','SINGULAR MATRIX, ELEMENT ',I2,I3,' DISEASE ',I1)
      IM=M-1
      DO 08320 I=1,IM
      IF (I.EQ.MQ) GO TO 08320
      A(I,I)=0.0
      T(I,I)=0.0
08320 CONTINUE
C
C     CALCULATE COEFICIENTS
09030 DO 09060 I=1,12
      C(I,K)=0.0
      DO 09060 J=1,12
09060 C(I,K)=T(I,J)*SIK(J,K)+C(I,K)
C
C     PRINT OUT COEFICIENTS
      WRITE (6,09140) (BT(K), K, (C(I,K),I=1,12))
09140 FORMAT ('0',F6.2,3X, I1, 6X, 12F9.4)
      F(K+6)=F(K)
      GO TO 07040
C
C     CALCULATE DISEASE-SYMPTOM FUNCTIONS OF PREVIOUS PATIENTS
13010 I=1
13020 CONTINUE
13030 FORMAT ('1',2X,'D',5X,'S1  S2  S3  S4  S5  S6  S7  S8  S9 S10 ',
     1   'S11', 18X,'D1', 8X,'D2', 8X,'D3', 8X,'D4', 8X,'D5', 8X,'D6')
```

```
13050 FORMAT(3X,I1,5X,11(I1,3X),I4, 6X, 6F10.4)
      WRITE (6,13030)
13070 DO 13110 K=1,6
      JOK=F(S(1,I)+6)
      D(K,I)=0.0
      DO 13100 J=1,12
13100 D(K,I)=D(K,I)+C(J,K)*S(J+1,I)
      D(K,I)=D(K,I)*F(K)/N
13110 DT(K,I)=D(K,I)*10.0
      WRITE(6,13050) (S(J,I),J=1,12),S(14,I),(D(K,I),K=1,6)
      JK=0
      DO 13150 J=1,6
      IF (J.EQ.S(1,I)) GO TO 13150
      IF (D(S(1,I),I).LT.D(J,I)) JK=1
13150 CONTINUE
      JCK=JOK-JK
      F(S(1,I)+6)=JOK
      IF (I.EQ.N) GO TO 13220
      I=I+1
      IF (S(1,I).EQ.S(1,I-1)) GO TO 13070
      GO TO 13020
13220 CONTINUE
C
C     CALCULATE DISEASE-SYMPTOM FUNCTIONS OF NEW PATIENTS
14030 MP=N
14040 MP=MP+1
      IF (MP.GT.NP) GO TO 14140
      IF (MP.EQ.N+1) WRITE(6,13030)
      DO 14110 K=1,6
      D(K,MP)=0.0
      DO 14100 J=1,12
14100 D(K,MP)=D(K,MP)+C(J,K)*S(J+1,MP)
14110 D(K,MP)=D(K,MP)*F(K)/N
      WRITE (6,13050) (S(J,MP),J=1,12),S(14,MP),(D(K,MP), K=1,6)
      GO TO 14040
14140 CONTINUE
C
C     CALCULATE DISEASE STATISTICS
      DO 16150 K1=1,6
      DO 16150 K=1,6
      QN(K,K1)=0
      RN(K,K1)=0
      DO 16140 I=1,12
      QN(K,K1)=QN(K,K1)+SIK(I,K)*C(I,K1)*F(K1)/(N*N)
      DO 16140 J=1,12
      RN(K,K1)=RN(K,K1)+C(I,K1)*AIJK(I,J,K)*C(J,K1)
16140 CONTINUE
16150 RN(K,K1)= F(K1)*SQRT(RN(K,K1)/N)/N
      DO 16260 J=1,6
      Q(J)=QN(J,J)
      R(J)=RN(J,J)
      Q(J+6)=0
      R(J+6)=0
      DO 16230 K=1,6
      IF (K.EQ.J) GO TO 16230
```

```
      Q(J+6)=Q(J+6)+QN(K,J)*F(K)
      R(J+6)=R(J+6)+RN(K,J)*F(K)*RN(K,J)
16230 CONTINUE
      Q(J+6)=Q(J+6)/(N-F(J))
      R(J+6)=SQRT(R(J+6)/(N-F(J)))
      E(J+6)=0.0
16260 IF(Q(J).NE.0.0) E(J+6)=R(J)/Q(J)
C
C     DETERMINE WEIGHTED MEAN SQUARE RATIO
      DO 16350 K=1,6
      WMSR(K)=0
      DO 16340 I=1,12
      DO 16330 J=1,12
16330 WMSR(K)=WMSR(K)+C(I,K)*SIJ(I,J)*C(J,K)
16340 CONTINUE
16350 IF(F(K).NE.0) WMSR(K)=(Q(K)*Q(K)*N*N/(WMSR(K)*F(K))-F(K)/N)
     1                              *(N/(N-F(K)))
C
C     PRINT OUT DISEASE STATISTICS
17030 FORMAT ('1   D    MEAN    STANDARD    RATIO    D    MEAN',
     1      '    STANDARD  NMS RATIO  POPULATION   CORRECT')
      WRITE (6,17030)
17050 FORMAT ('0',I4,F09.4,F10.4,F10.4,5X,'N',I1,F10.4,F10.4,
     1          F10.4,1X,       F10.1,3X,F10.1)
      DO 17070 J=1,6
17070 WRITE (6,17050) J,Q(J),R(J),E(J+6),J,Q(J+6),R(J+6),WMSR(J),
     1   F(J),F(6+J)
      F(13)=N
      F(14)=0.0
      F(15)=0.0
      DO 17125 J=1,6
      F(15)=F(15)+WMSR(J)/KD
      F(14)=F(14)+F(J+6)
17125 F(J+6)=F(J)
17130 FORMAT ('0', 60X,F10.4,1X,F10.1,3X,F10.1)
      WRITE (6,17130) F(15), F(13), F(14)
      DO 17160 I=1,12
      ENT(I)=0.0
17160 PAVG(I)=0.0
C
C     CALCULATE EXTENDED PROBABILITIES OF PREVIOUS AND NEW PATIENTS
      ID=1
18030 WRITE(6,13030)
18040 JOK=F(S(1,ID)+6)
      DO 18190 K1=1,KD
      P(K1+6)=0.0
      DO 18130 K=1,KD
      IF (K1.EQ.K)GO TO 18130
      Y(1)=((D(K1,ID)-QN(K,K1))/(1.414*RN(K,K1)))**2
      IF(Y(1).GT.174.0) Y(1)=174.0
      P(K1+6)=P(K1+6)+F(K)/(EXP(Y(1))*RN(K,K1))
18130 CONTINUE
      Y(1)=((D(K1,ID)-Q(K1))/(1.414*R(K1)))**2
      IF(Y(1).GT.174.0) Y(1)=174.0
      P(K1)=F(K1)/(EXP(Y(1))*R(K1))
```

```
      P(K1+12)=P(K1)/(P(K1)+P(K1+6))
      IF (S(1,ID).EQ.K1) PAVG(K1)=PAVG(K1)+P(K1+12)/F(K1)
      IF (S(1,ID).EQ.K1) ENT(K1)=ENT(K1)-ALOG(P(K1+12))/F(K1)
      IF (S(1,ID).NE.K1) PAVG(K1+6)=PAVG(K1+6)+P(K1+12)/(N-F(K1))
      IF (S(1,ID).NE.K1) ENT(K1)=ENT(K1)-ALOG(1.0-P(K1+12))/(N-F(K1))
18190 CONTINUE
      WRITE(6,13050) (S(J,ID),J=1,12),S(14,ID),(P(J+12), J=1,6)
      JK=0
      DO 18205 J=1,6
      IF (J.EQ.S(1,ID)) GO TO 18205
      IF (P(S(1,ID)+12).LT.P(J+12)) JK=1
18205 CONTINUE
      JCK=JOK-JK
      F(S(1,ID)+6)=JOK
      ID=ID+1
      IF (ID.LE.N) GO TO 18260
      IF (ID.GT.NP) GO TO 18280
      IF (ID.EQ.N+1) WRITE(6,13030)
      GO TO 18040
18260 IF (S(1,ID).NE.S(1,ID-1)) GO TO 18030
      GO TO 18040
18280 CONTINUE
C
C     CALCULATE FURTHER EXTENDED PROBABILITIES OF NEW PATIENTS
      ID=N+1
18330 IF (ID.GT.NP) GO TO 18580
      IF (ID.EQ.N+1) WRITE(6,13030)
18340 DO 18430 K =1,KD
      P(K+6)=F(K)/N
      DO 18430 K1=1,KD
      Y(K1)=((D(K1,ID)-QN(K,K1))/(1.414*RN(K,K1)))**2
      IF(Y(K1).GT.174.0) Y(K1)=174.0
      P(K+6)=P(K+6)*1.0/(EXP(Y(K1))*RN(K,K1))
18430 CONTINUE
      P(19)=0
      DO 18470 K=1,6
18470 P(19)=P(19)+P(K+6)
      DO 18490 K=1,6
18490 P(K+12)=P(K+6)/P(19)
      WRITE(6,13050) (S(J,ID),J=1,12),S(14,ID),(P(J+12), J=1,6)
      ENT(S(1,ID)+6)=ENT(S(1,ID)+6)-ALOG(P(S(1,ID)+12))/F(S(1,ID))
      ID=ID+1
      GO TO 18330
18580 CONTINUE
C
C     CALCULATE PROBABILITY STATISTICS
      I=0
19040 DO 19140 K1=1,KD
      P(K1+6)=0
      DO 19110 K=1,KD
      IF (K1.EQ.K) GO TO 19110
      Y(1)=((Q(K1+I)-QN(K,K1))/(1.414*RN(K,K1)))**2
      IF (Y(1).GT.174.0) Y(1)=174.0
      P(K1+6)=P(K1+6)+F(K)/(EXP(Y(1))*RN(K,K1))
19110 Y(1)=((Q(K1+I)-Q(K1))/(1.414*R(K1)))**2
```

```
      IF (Y(1).GT.174.0) Y(1)=174.0
      P(K1)=F(K1)/(EXP(Y(1))*R(K1))
19140 PMEAN(K1+I)=P(K1)/(P(K1)+P(K1+6))
      I=I+6
      IF (I.EQ.6) GO TO 19040
      DO 19190 I=1,KD
      PAVG(I+12)=PAVG(I)*(1-PAVG(I+6))/((1-PAVG(I))*PAVG(I+6))
19190 Y(I)=ALOG(PMEAN(I)*(1.0-PMEAN(I+6))/((1.0-PMEAN(I))*PMEAN(I+6)))
C
C     PRINT OUT PROBABILITY STATISTICS
19220 FORMAT('1   D    MEAN     D    MEAN    PRODUCT   D  AVERAGE   ',
     1    'D  AVERAGE   PRODUCT   ENTROPY  POPULATION    CORRECT')
19250 FORMAT ('0',I4,F9.4,3X,'N',I1,F9.4,F10.4,I4,F9.4,3X,'N',
     1   I1,F9.4,F10.4,F10.4,F10.1,3X,F10.1)
      WRITE(6,19220)
      DO 19280 J=1,KD
19280 WRITE(6,19250) (J,PMEAN(J),J,PMEAN(J+6),Y(J),J,
     1    PAVG(J),J,PAVG(J+6),PAVG(J+12),ENT(J),F(J),F(J+6))
      F(13)=N
      F(14)=F(7)
      DO 19330 I=2,6
      ENT(1)=ENT(1)+ENT(I)
19330 F(14)=F(14)+F(I+6)
      ENT(1)=ENT(I)/KD
19340 FORMAT('0',74X,F10.4,F10.1,3X,F10.1)
      WRITE (6,19340) (ENT(1),F(13),F(14))
C
C     CONTROL LOOP FOR INCREMENTS IN BETA
      DO 20040 I=1,6
20040 BT(I)=BT(I)+BTINCR
      IF(BT(1).LT.0.4) GO TO 07030
C
C     CONTROL LOOP FOR Q'TH SYMPTOM
      MQ=MQ-1
      IF (MQ.EQ.0) GO TO 20130
      GO TO 06300
C
C     CONTROL LOOP FOR NUMBER OF SYMPTOMS USED
20130 M=M-1
      MQ=M-1
      IF (M.NE.1) GO TO 06300
20150 CONTINUE
C
      END
```